Data-driven Characterization of Fake Content in Social Media

FAN Cheuk Pan

Agenda

- Background
- Methodology
- Data Acquisition
- Classification Model
- Application Implementation
- Conclusion

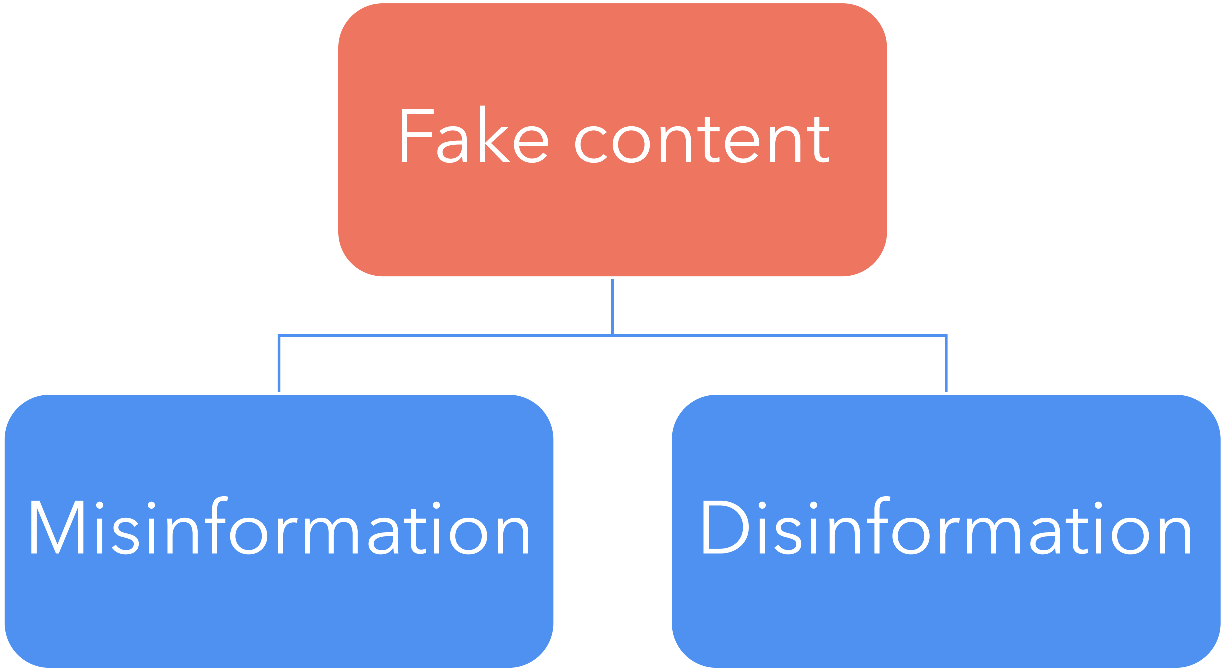
Background

Background

Problem Statement

Objectives

Background



Fake content

Misinformation

Disinformation

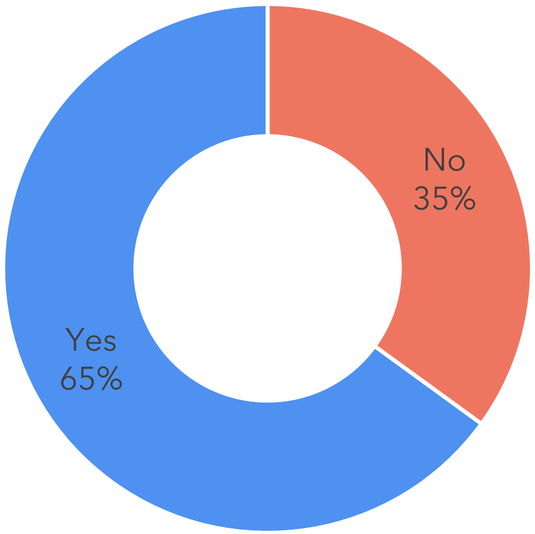
Background

2019 CIGI-Ipsos Global Survey on

Internet Security and Trust

CIGI Ipsos Global Survey 2019

Reported seeing fake news on social media



Problem Statement

Cause economic loss	• Fake shop
Gain advantage by bad intention	• Gain vote in election by spread disinformation
Cause social panic	• False COVID-19 treatments • Panic Buying

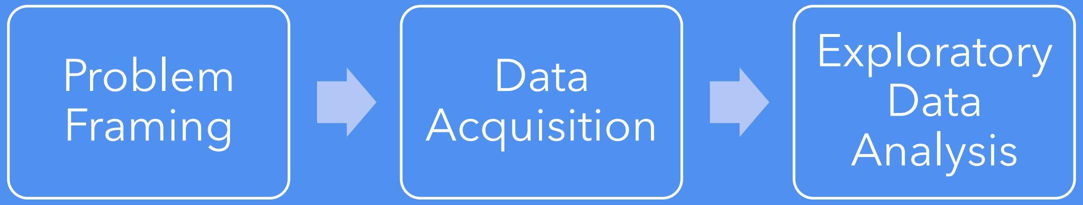
Objectives

Verify fake content

Analysis of fake content

Rising awareness of fake content

Methodology

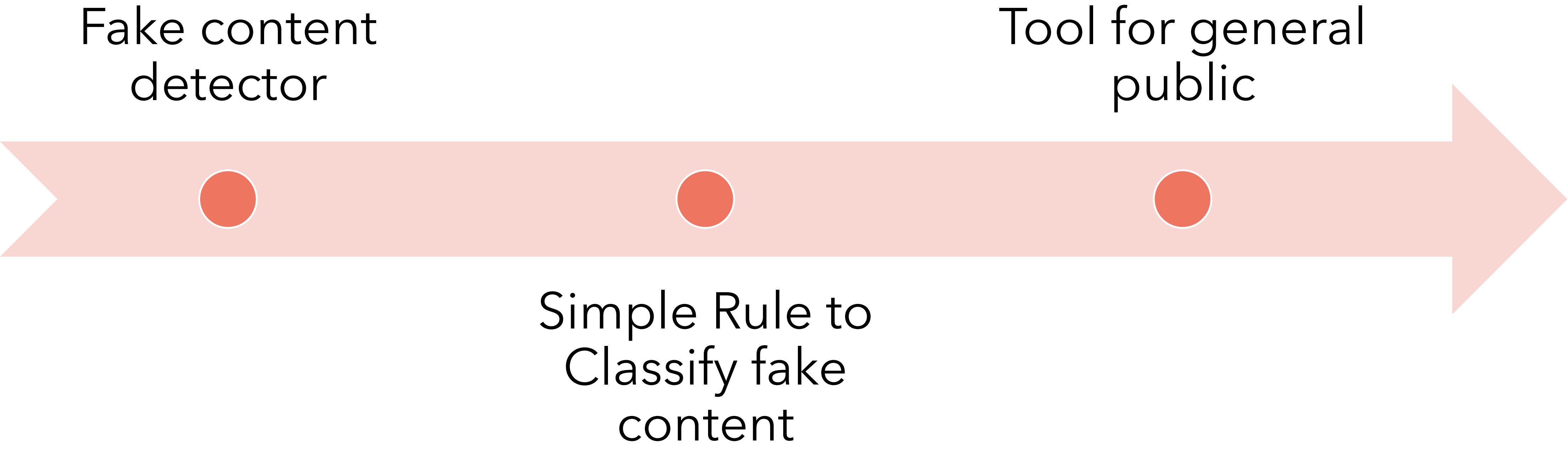


Problem Framing

Data Acquisition

Exploratory Data Analysis

Methodology

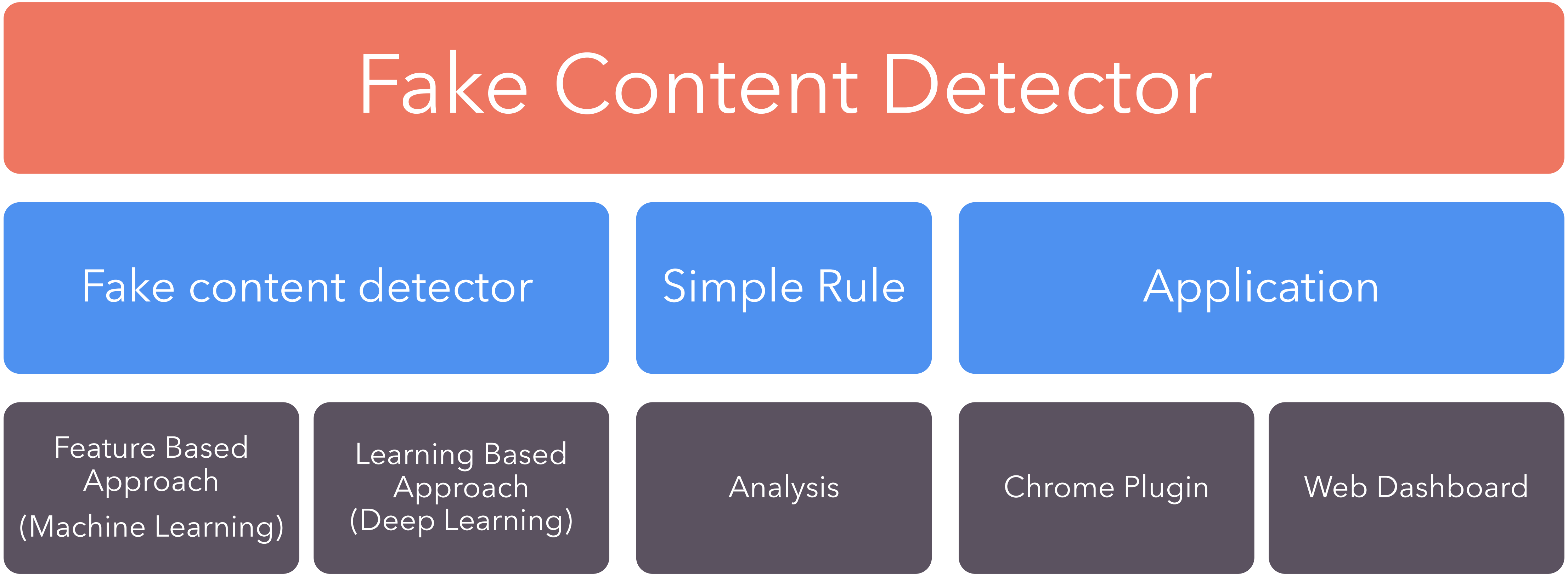


Fake content detector

Simple Rule to Classify fake content

Tool for general public

System Component



Fake Content Detector

Fake content detector

Simple Rule

Application

Feature Based Approach (Machine Learning)

Learning Based Approach (Deep Learning)

Analysis

Chrome Plugin

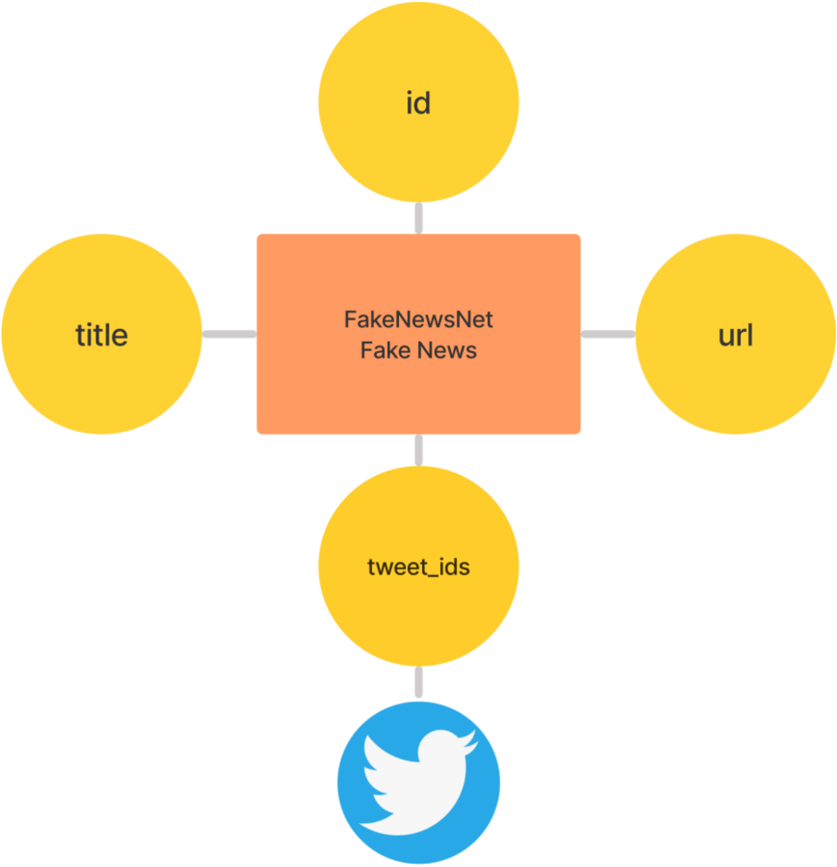
Web Dashboard

Data Acquisition

Data Acquisition – Data Source

Data Source

- FakeNewsNet
  - Gossip Cop
  - PolitiFact
- Twitter Data



Data Acquisition – Fetch from Twitter

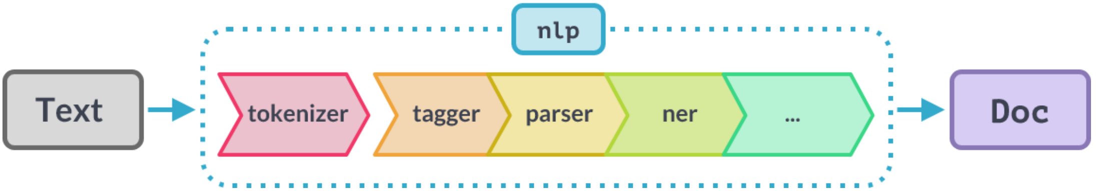
Tweet
Content
Public metrics
Hashtag
URL
Context Domain
Created Time
Label

User
Public metrics
Description
Created time

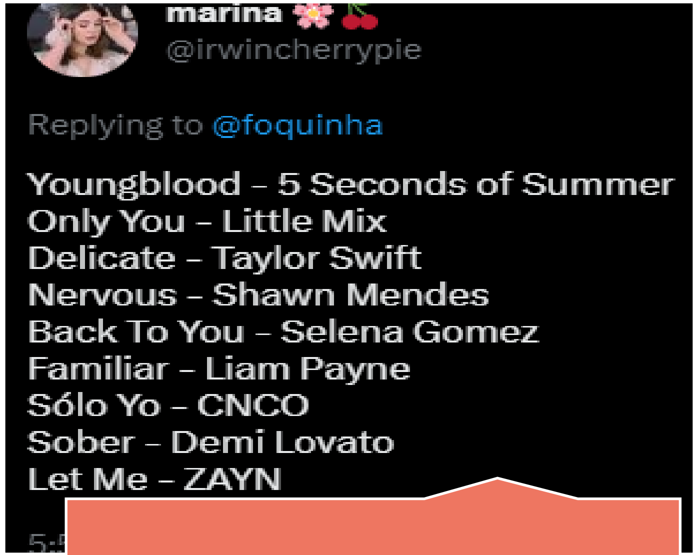
Data Acquisition – Pre-Processing

Python Spacy

- Tokenization
- Remove
  - Stop word
  - Number
- Lemmatization
- Lower case



Data Acquisition – Pre-Processing



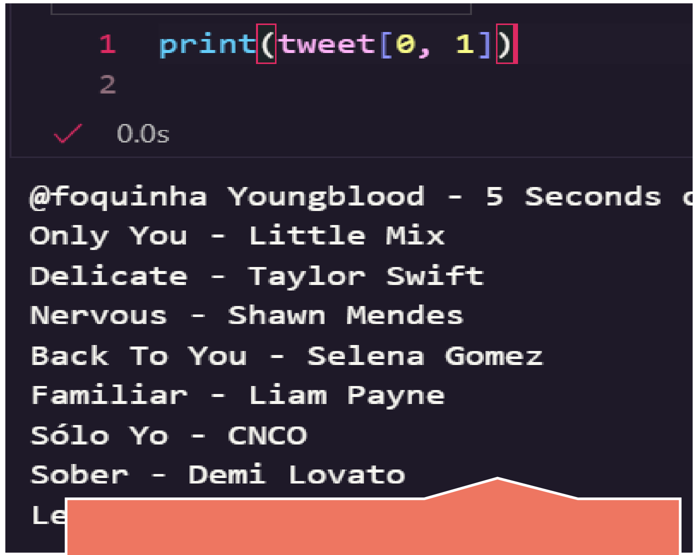
marina 🌸🍒
@irwincherrypie

Replying to @foquinha

Youngblood – 5 Seconds of Summer
Only You – Little Mix
Delicate – Taylor Swift
Nervous – Shawn Mendes
Back To You – Selena Gomez
Familiar – Liam Payne
Sólo Yo – CNCO
Sober – Demi Lovato
Let Me – ZAYN

5:5

Data Source

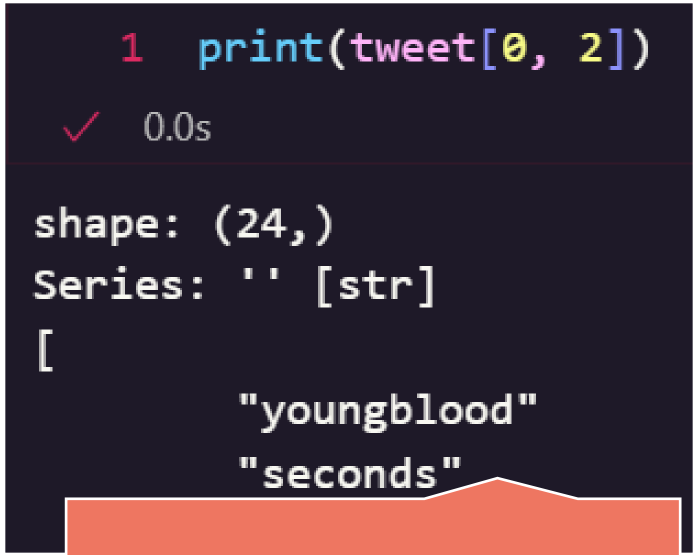


```
print(tweet[0, 1])
```

✓ 0.0s

```
@foquinha Youngblood - 5 Seconds c
Only You - Little Mix
Delicate - Taylor Swift
Nervous - Shawn Mendes
Back To You - Selena Gomez
Familiar - Liam Payne
Sólo Yo - CNCO
Sober - Demi Lovato
Le
```

Fetch from data



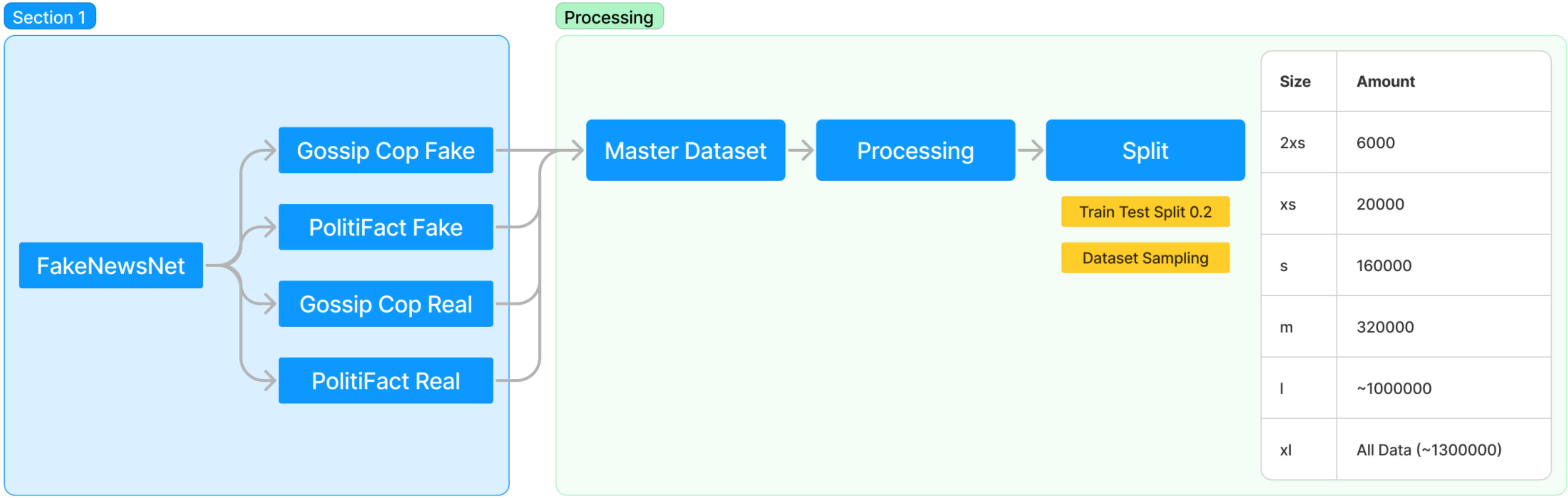
```
print(tweet[0, 2])
```

✓ 0.0s

```
shape: (24,)
Series: '' [str]
[
        "youngblood"
        "seconds"
```

Processed Text

Data Acquisition – Split Data

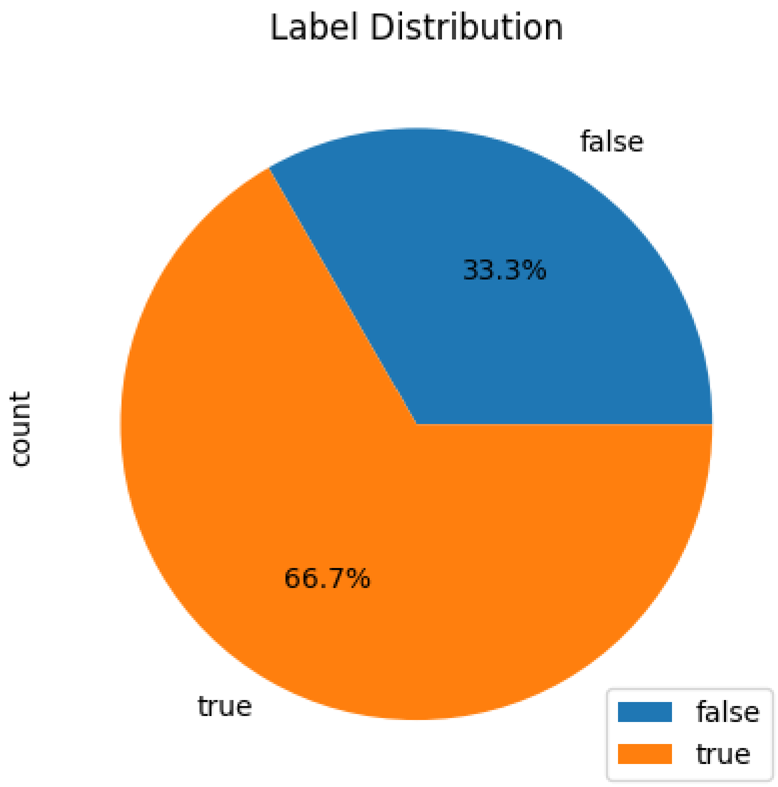


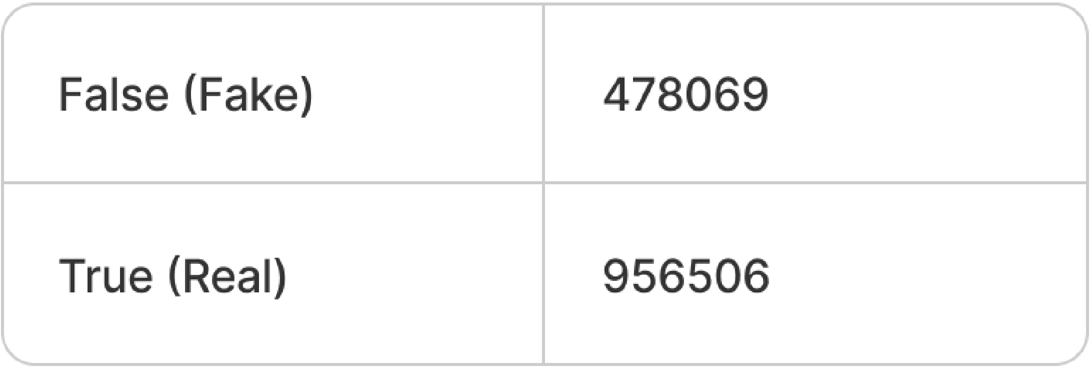
Size	Amount
2xs	6000
xs	20000
s	160000
m	320000
l	~1000000
xl	All Data (~1300000)

Exploratory Data Analysis (EDA)

Total Record	• Label Distribution
Sentence Length	• Word • Character
Term Frequency	• Word Cloud • Word Ranking
Meta data	• Hashtag • Domain (Annotated by Twitter)

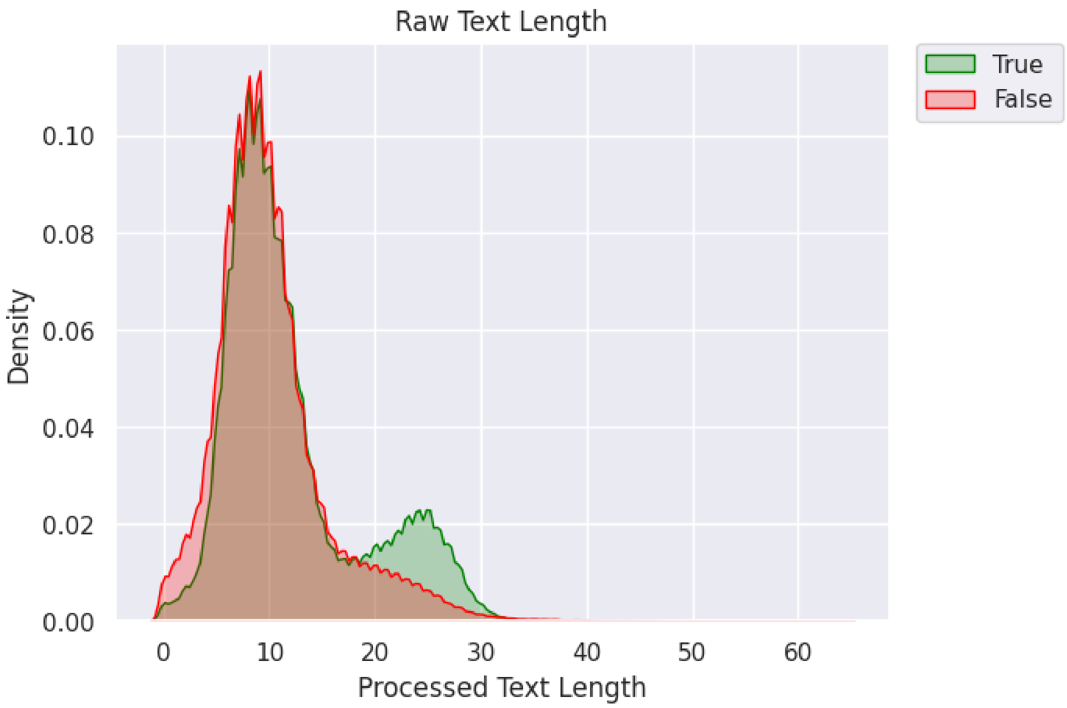
EDA – Label ExDistribution

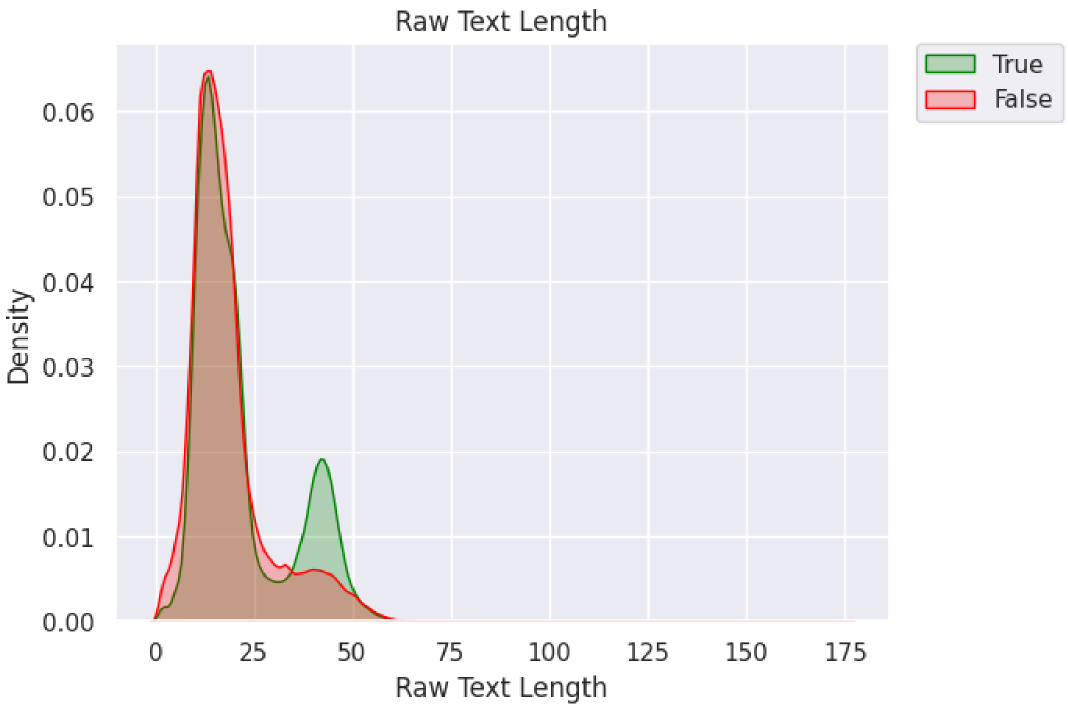




False (Fake)	478069
True (Real)	956506

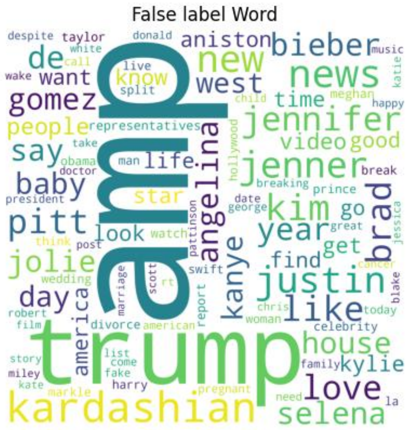
EDA – Sentence Length



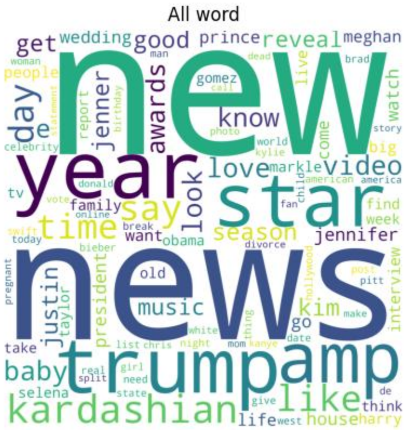


EDA – Word Cloud of Content

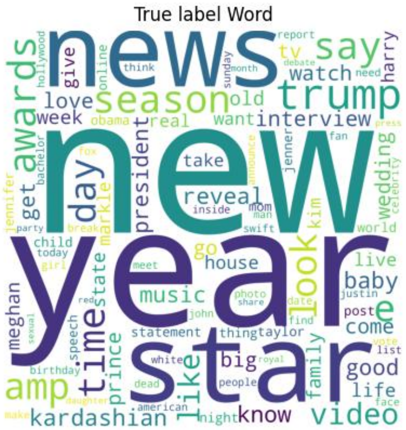
False Label



All

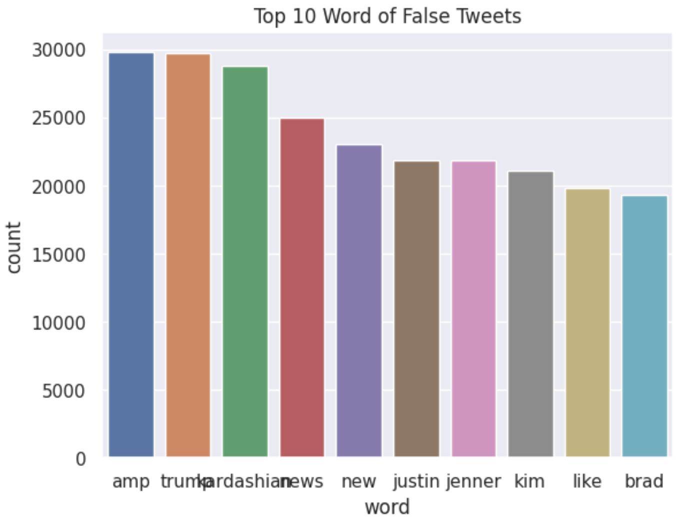


True Label

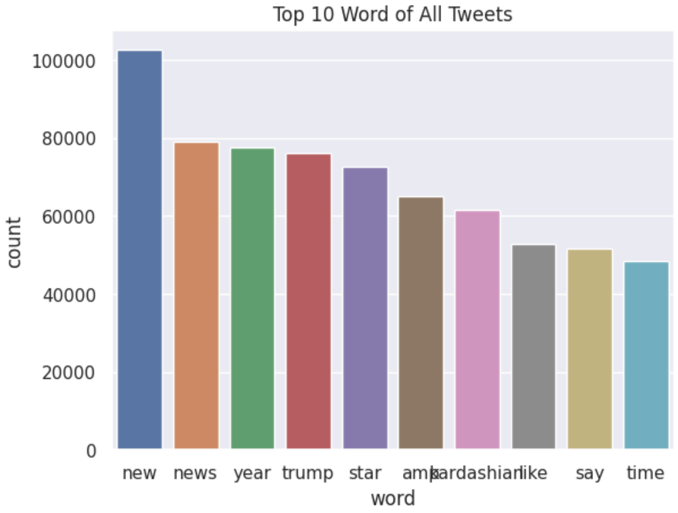


EDA – Word Cloud of Content

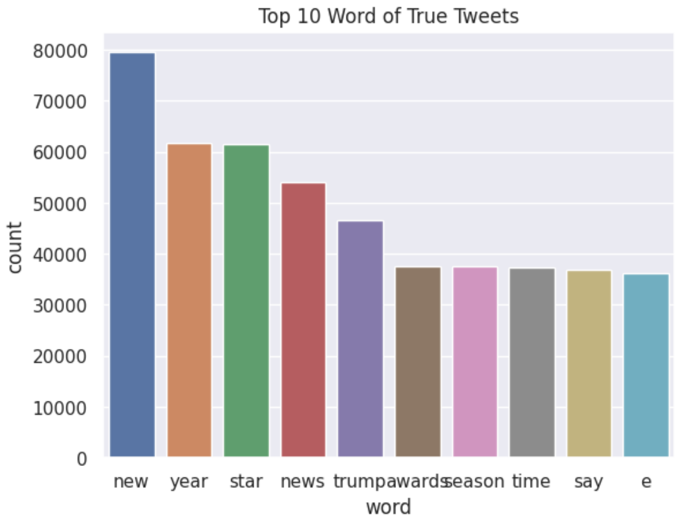
False Label



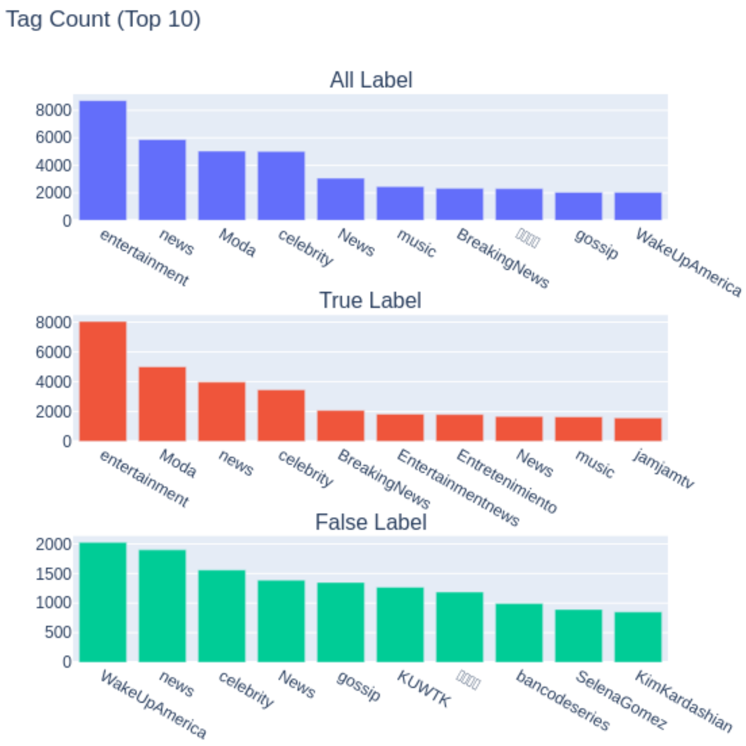
All



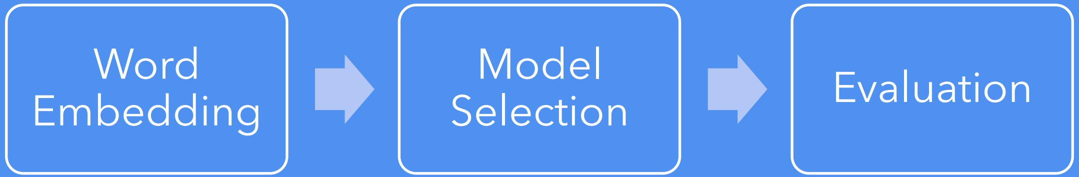
True Label



EDA – Top 10 Hashtags



Classification Model

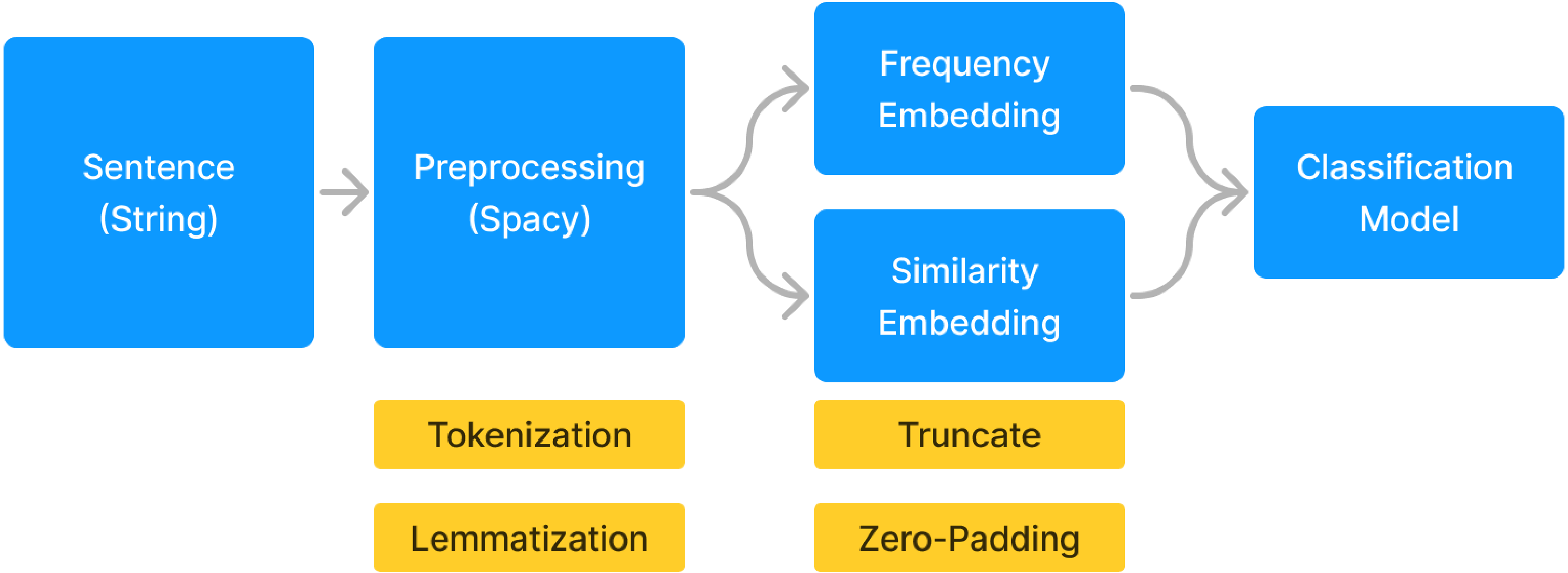


Word Embedding

Model Selection

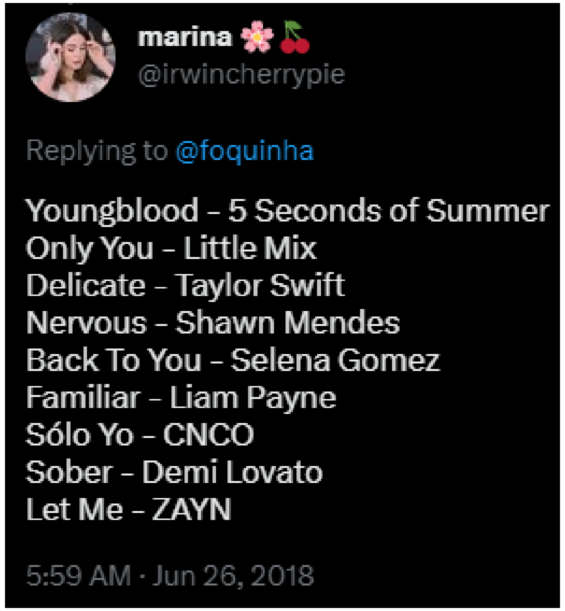
Evaluation

Classification Model – Procedure



Word Embedding – Frequency

- Count (BOW)
- TF-IDF

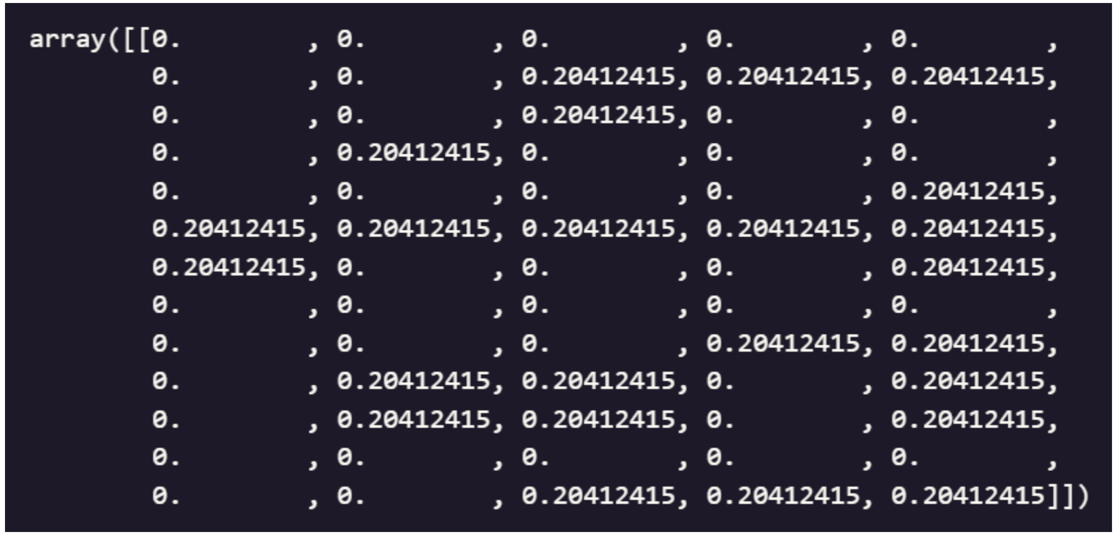


marina 🌸🍒
@irwincherrypie

Replying to @foquinha

Youngblood – 5 Seconds of Summer
Only You – Little Mix
Delicate – Taylor Swift
Nervous – Shawn Mendes
Back To You – Selena Gomez
Familiar – Liam Payne
Sólo Yo – CNCO
Sober – Demi Lovato
Let Me – ZAYN

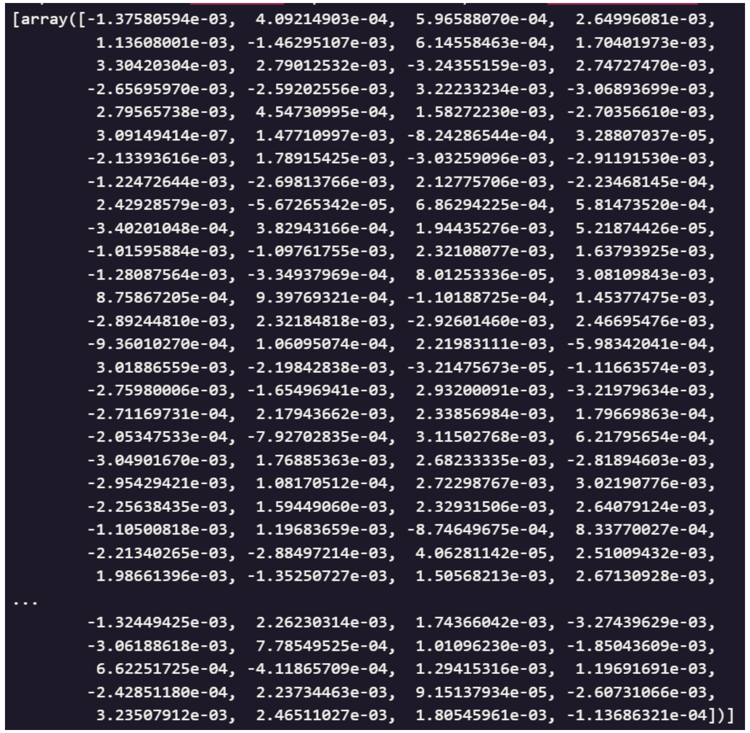
5:59 AM · Jun 26, 2018



```
array([[0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.20412415, 0.20412415, 0.20412415,
        0.        , 0.        , 0.20412415, 0.        , 0.        ,
        0.        , 0.20412415, 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.20412415,
        0.20412415, 0.20412415, 0.20412415, 0.20412415, 0.20412415,
        0.20412415, 0.        , 0.        , 0.        , 0.20412415,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.20412415, 0.20412415,
        0.        , 0.20412415, 0.20412415, 0.        , 0.20412415,
        0.        , 0.20412415, 0.20412415, 0.        , 0.20412415,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.20412415, 0.20412415, 0.20412415]])
```

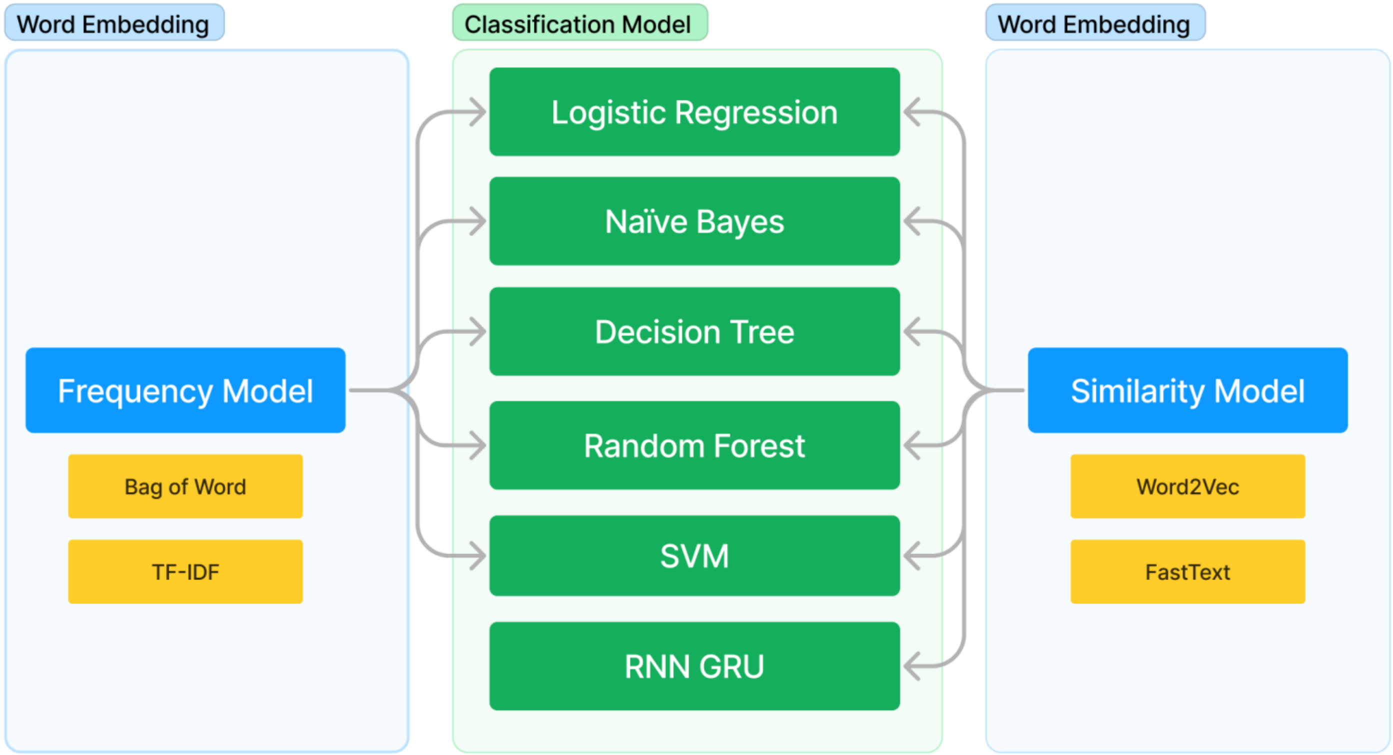
Word Embedding – Similarity

- Custom Train from Pre-trained
  - Word2Vec
    - Score: 0.450461
  - FastText
    - Score: 0.226015



```
[array([-1.37580594e-03,  4.09214903e-04,  5.96588070e-04,  2.64996081e-03,
         1.13608001e-03, -1.46295107e-03,  6.14558463e-04,  1.70401973e-03,
         3.30420304e-03,  2.79012532e-03, -3.24355159e-03,  2.74727470e-03,
        -2.65695970e-03, -2.59202556e-03,  3.22233234e-03, -3.06893699e-03,
         2.79565738e-03,  4.54730995e-04,  1.58272230e-03, -2.70356610e-03,
         3.09149414e-07,  1.47710997e-03, -8.24286544e-04,  3.28807037e-05,
        -2.13393616e-03,  1.78915425e-03, -3.03259096e-03, -2.91191530e-03,
        -1.22472644e-03, -2.69813766e-03,  2.12775706e-03, -2.23468145e-04,
         2.42928579e-03, -5.67265342e-05,  6.86294225e-04,  5.81473520e-04,
        -3.40201048e-04,  3.82943166e-04,  1.94435276e-03,  5.21874426e-05,
        -1.01595884e-03, -1.09761755e-03,  2.32108077e-03,  1.63793925e-03,
        -1.28087564e-03, -3.34937969e-04,  8.01253336e-05,  3.08109843e-03,
         8.75867205e-04,  9.39769321e-04, -1.10188725e-04,  1.45377475e-03,
        -2.89244810e-03,  2.32184818e-03, -2.92601460e-03,  2.46695476e-03,
        -9.36010270e-04,  1.06095074e-04,  2.21983111e-03, -5.98342041e-04,
         3.01886559e-03, -2.19842838e-03, -3.21475673e-05, -1.11663574e-03,
        -2.75980006e-03, -1.65496941e-03,  2.93200091e-03, -3.21979634e-03,
        -2.71169731e-04,  2.17943662e-03,  2.33856984e-03,  1.79669863e-04,
        -2.05347533e-04, -7.92702835e-04,  3.11502768e-03,  6.21795654e-04,
        -3.04901670e-03,  1.76885363e-03,  2.68233335e-03, -2.81894603e-03,
        -2.95429421e-03,  1.08170512e-04,  2.72298767e-03,  3.02190776e-03,
        -2.25638435e-03,  1.59449060e-03,  2.32931506e-03,  2.64079124e-03,
        -1.10500818e-03,  1.19683659e-03, -8.74649675e-04,  8.33770027e-04,
        -2.21340265e-03, -2.88497214e-03,  4.06281142e-05,  2.51009432e-03,
         1.98661396e-03, -1.35250727e-03,  1.50568213e-03,  2.67130928e-03,
...
        -1.32449425e-03,  2.26230314e-03,  1.74366042e-03, -3.27439629e-03,
        -3.06188618e-03,  7.78549525e-04,  1.01096230e-03, -1.85043609e-03,
         6.62251725e-04, -4.11865709e-04,  1.29415316e-03,  1.19691691e-03,
        -2.42851180e-04,  2.23734463e-03,  9.15137934e-05, -2.60731066e-03,
         3.23507912e-03,  2.46511027e-03,  1.80545961e-03, -1.13686321e-04])]
```

Classification Model



Word Embedding

Frequency Model

Bag of Word

TF-IDF

Classification Model

Logistic Regression

Naïve Bayes

Decision Tree

Random Forest

SVM

RNN GRU

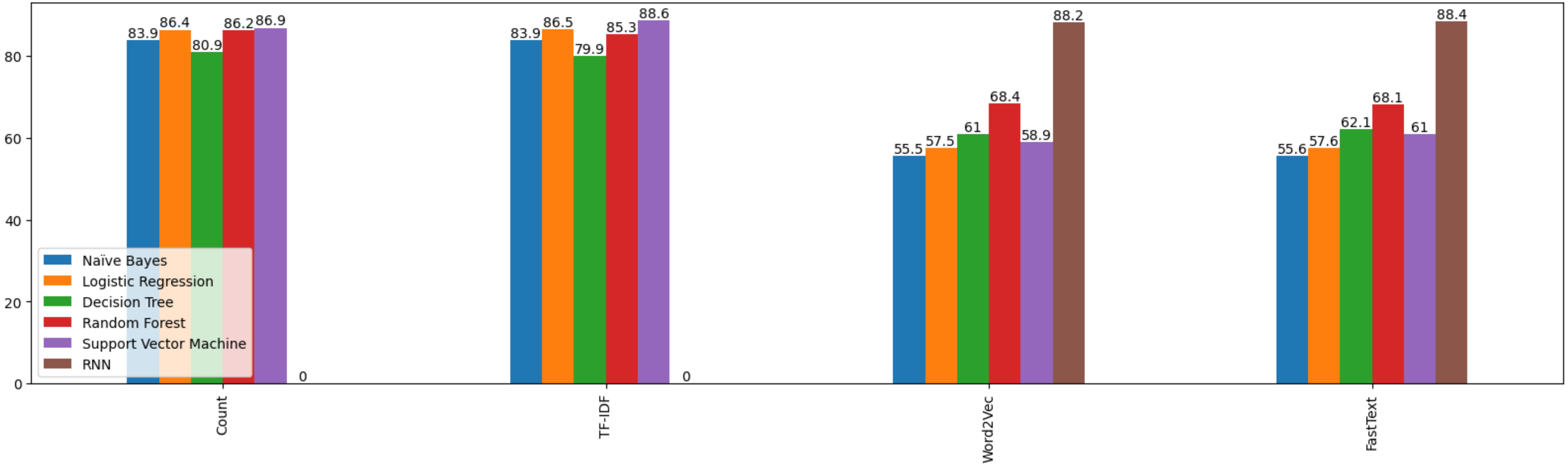
Word Embedding

Similarity Model

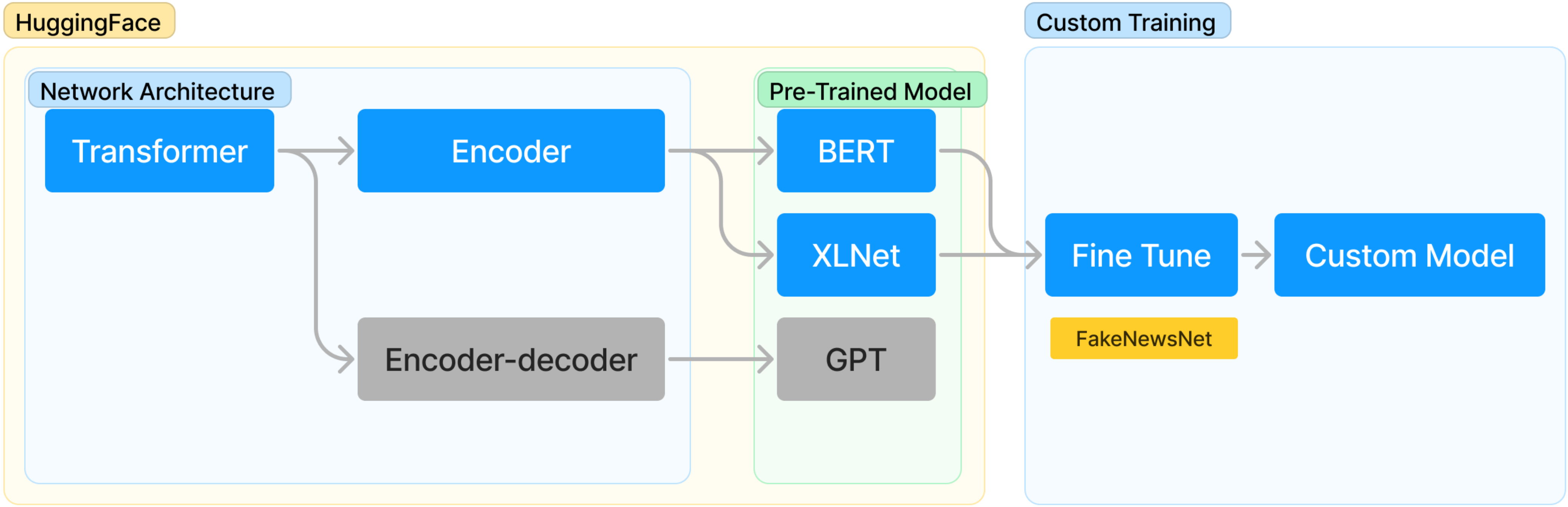
Word2Vec

FastText

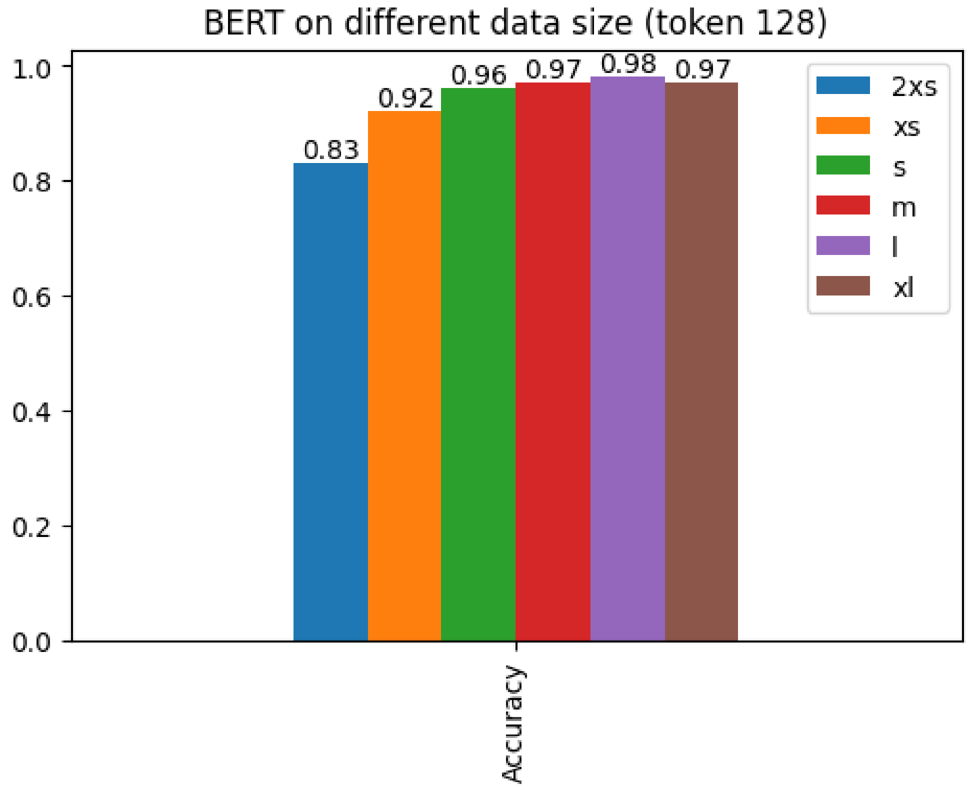
Classification Model – Result (Accuracy)



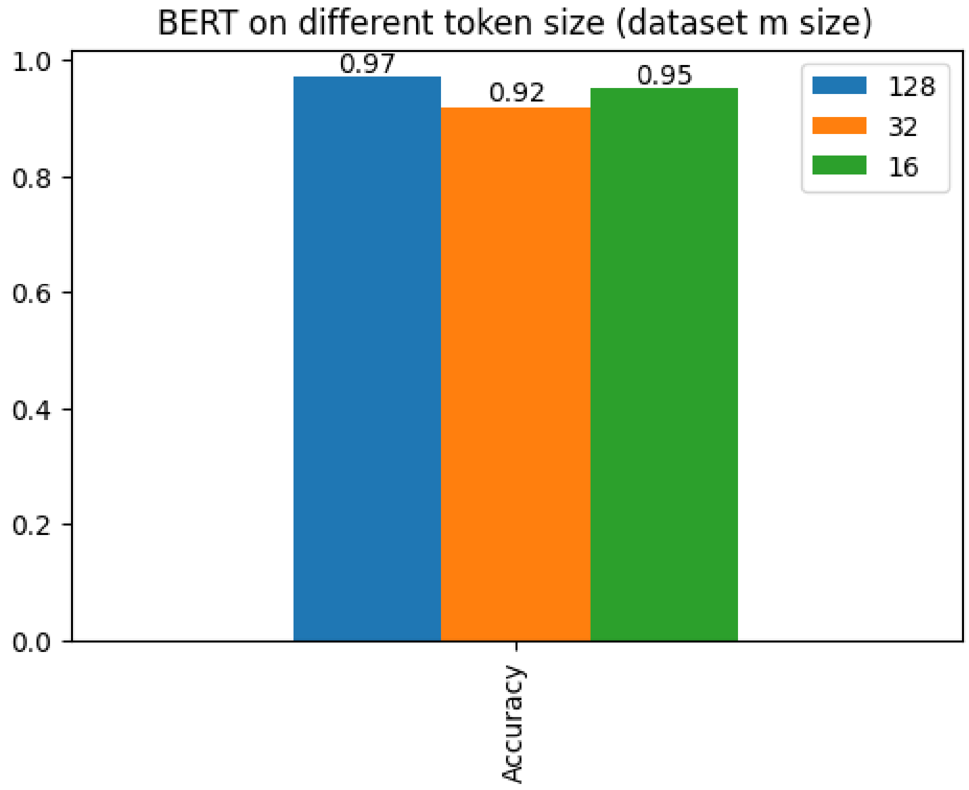
Transformer Model



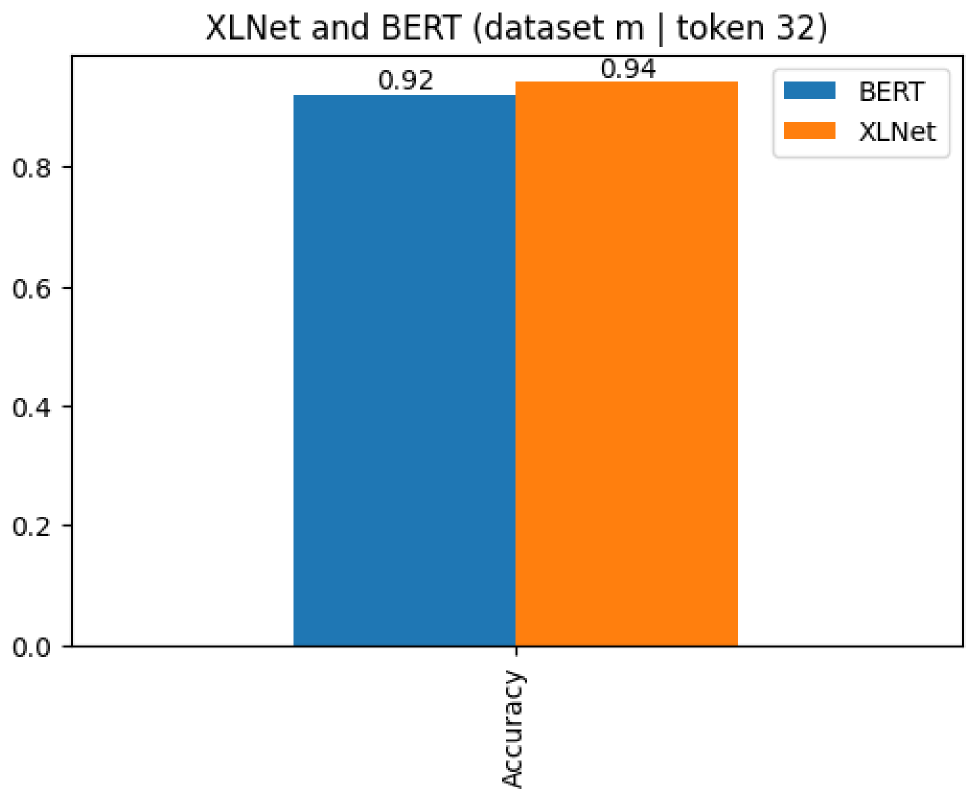
Transformer Model – BERT on different data size



Transformer Model – BERT on different token size



Transformer Model – BERT and XLNet



Model – Result conclusion

Feature-based

- Achieve more than 80% accuracy
- Neural Network work well with Similarity Word Embedding Model

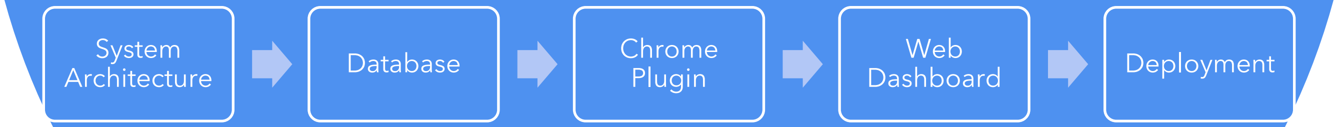
Learning-based

- Achieve over 90% accuracy
- More data result in more
- Short-text classification, smaller token size not sightly affect the result
- Same training data size, XLNet may perform better

Overfit

- No limitation on maximum feature

Application



System Architecture → Database → Chrome Plugin → Web Dashboard → Deployment

Recall Objectives

Verify fake content

- Fake content detector

Analysis of fake content

- Simple Rule to Classify fake content

Rising awareness of fake content

- Tool for general public

Application

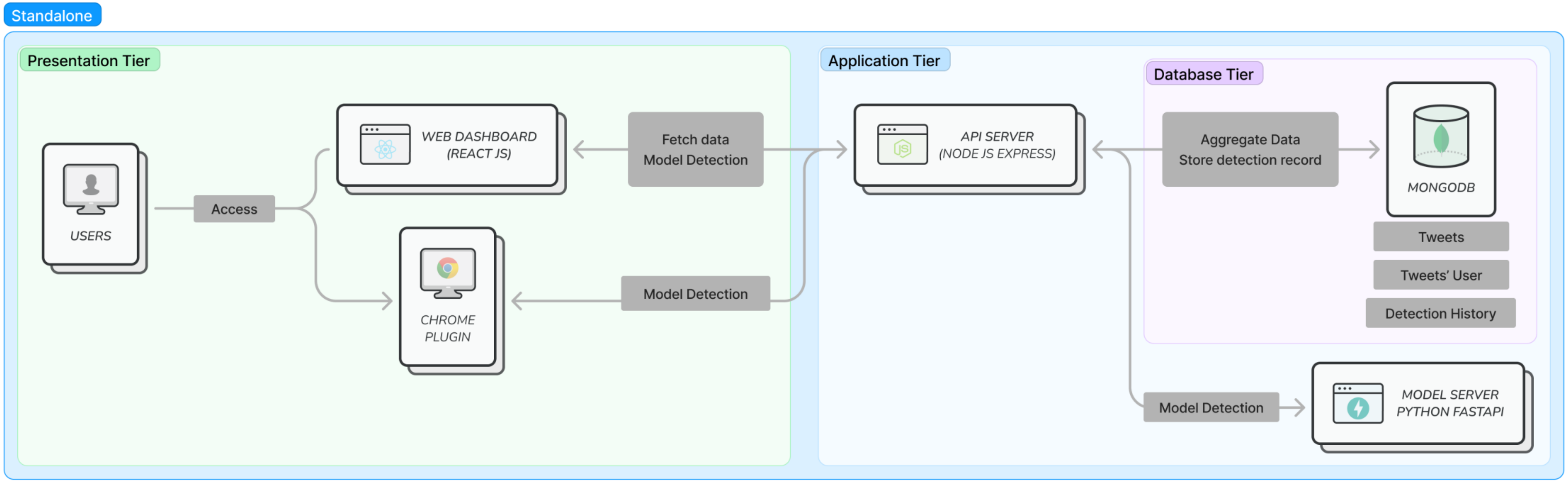
Chrome Plugin

- Perform detection

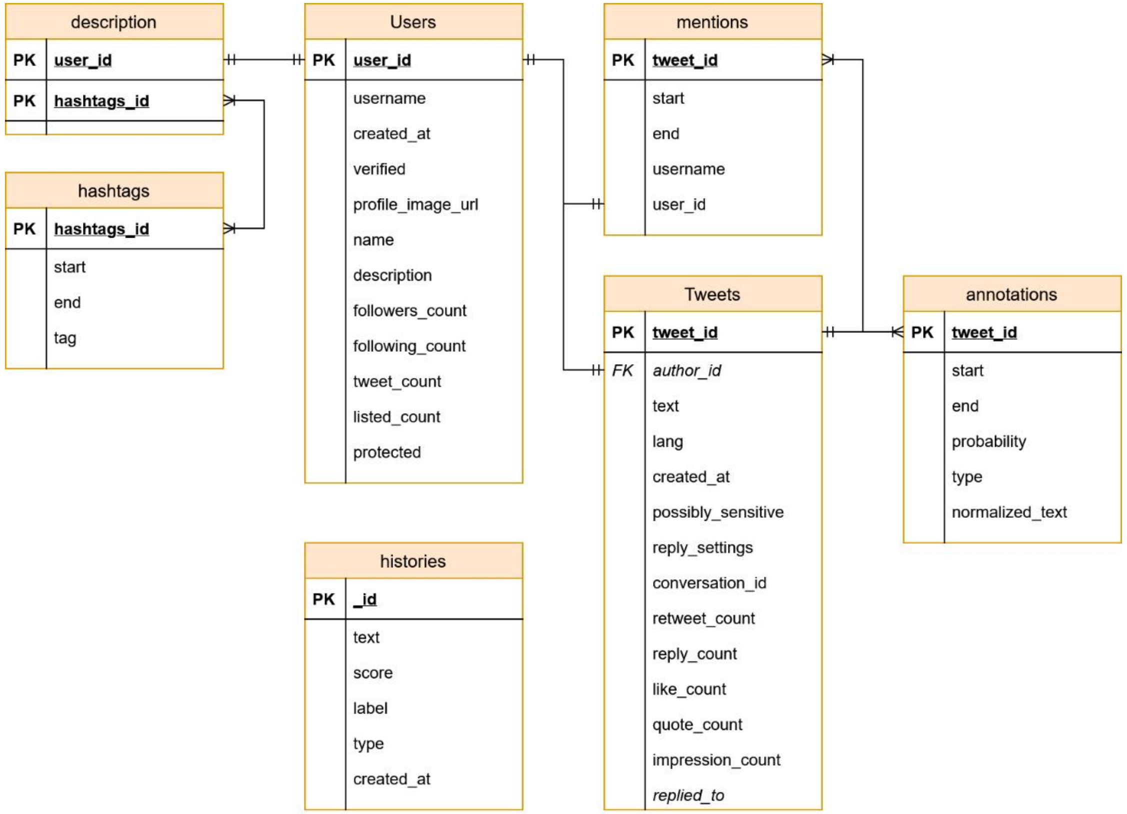
Web Dashboard

- Perform detection
- Data Visualization of dataset

System Architecture



Database



description	
PK	**user_id**
PK	**hashtags_id**

hashtags	
PK	**hashtags_id**
	start
	end
	tag

Users	
PK	**user_id**
	username
	created_at
	verified
	profile_image_url
	name
	description
	followers_count
	following_count
	tweet_count
	listed_count
	protected

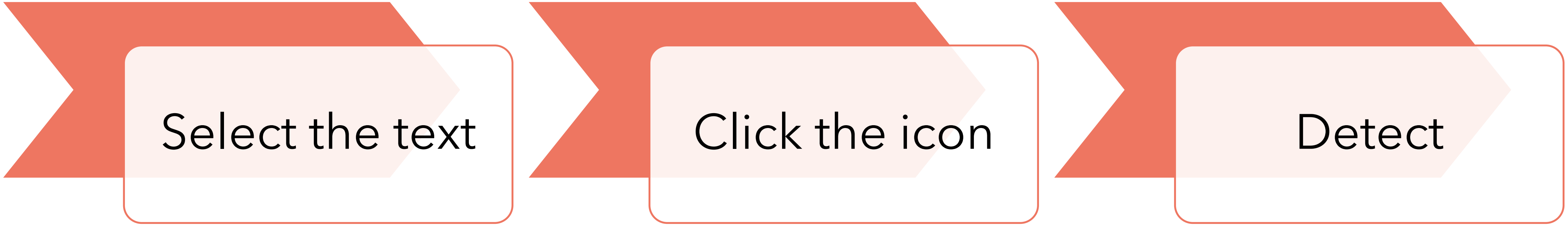
histories	
PK	**_id**
	text
	score
	label
	type
	created_at

mentions	
PK	**tweet_id**
	start
	end
	username
	user_id

Tweets	
PK	**tweet_id**
FK	*author_id*
	text
	lang
	created_at
	possibly_sensitive
	reply_settings
	conversation_id
	retweet_count
	reply_count
	like_count
	quote_count
	impression_count
	replied_to

annotations	
PK	**tweet_id**
	start
	end
	probability
	type
	normalized_text

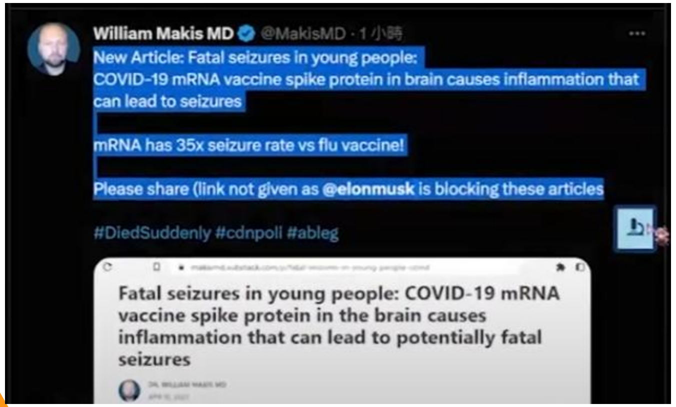
Chrome Plugin



Select the text

Click the icon

Detect



William Makis MD @MakisMD · 1 小時

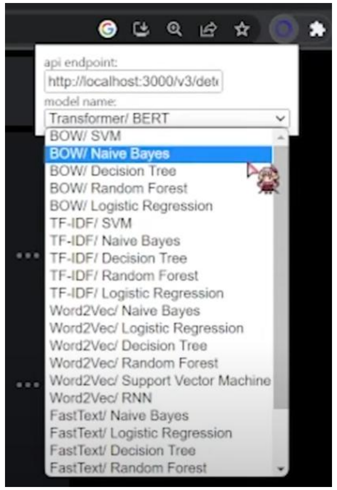
New Article: Fatal seizures in young people:
COVID-19 mRNA vaccine spike protein in brain causes inflammation that can lead to seizures

mRNA has 35x seizure rate vs flu vaccine!

Please share (link not given as @elonmusk is blocking these articles

#DiedSuddenly #cdnpoli #ableg

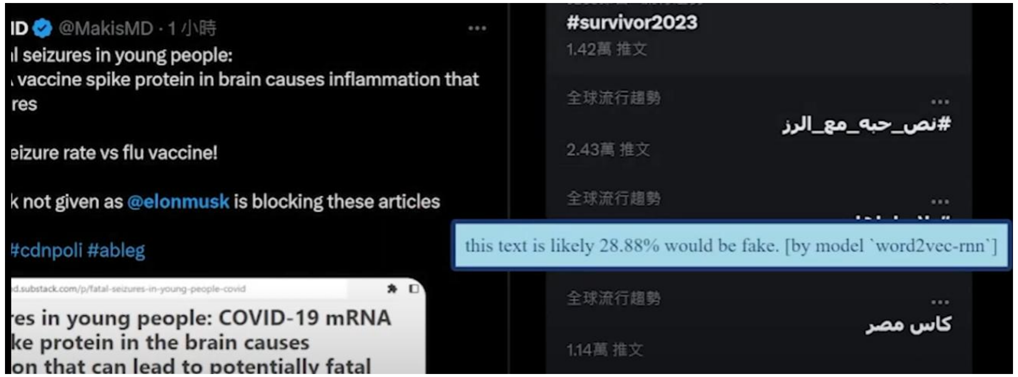
Fatal seizures in young people: COVID-19 mRNA vaccine spike protein in the brain causes inflammation that can lead to potentially fatal seizures



api endpoint: http://localhost:3000/v3/det

model name: Transformer/ BERT

BOW/ SVM
BOW/ Naive Bayes
BOW/ Decision Tree
BOW/ Random Forest
BOW/ Logistic Regression
TF-IDF/ SVM
TF-IDF/ Naive Bayes
TF-IDF/ Decision Tree
TF-IDF/ Random Forest
TF-IDF/ Logistic Regression
Word2Vec/ Naive Bayes
Word2Vec/ Logistic Regression
Word2Vec/ Decision Tree
Word2Vec/ Random Forest
Word2Vec/ Support Vector Machine
Word2Vec/ RNN
FastText/ Naive Bayes
FastText/ Logistic Regression
FastText/ Decision Tree
FastText/ Random Forest



ID @MakisMD · 1 小時

l seizures in young people:
vaccine spike protein in brain causes inflammation that
res

eizure rate vs flu vaccine!

k not given as @elonmusk is blocking these articles

#cdnpoli #ableg

d.substack.com/p/fatal-seizures-in-young-people-covid

es in young people: COVID-19 mRNA
ke protein in the brain causes
on that can lead to potentially fatal

#survivor2023
1.42萬 推文

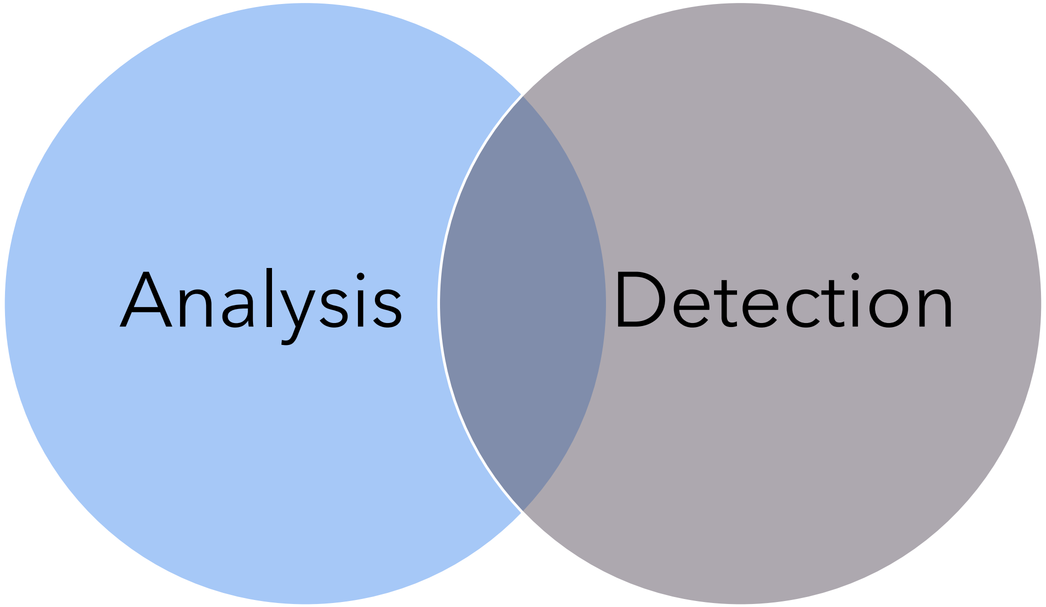
全球流行趨勢
#نص_حبه_مع_الرز
2.43萬 推文

全球流行趨勢

this text is likely 28.88% would be fake. [by model `word2vec-rnn`]

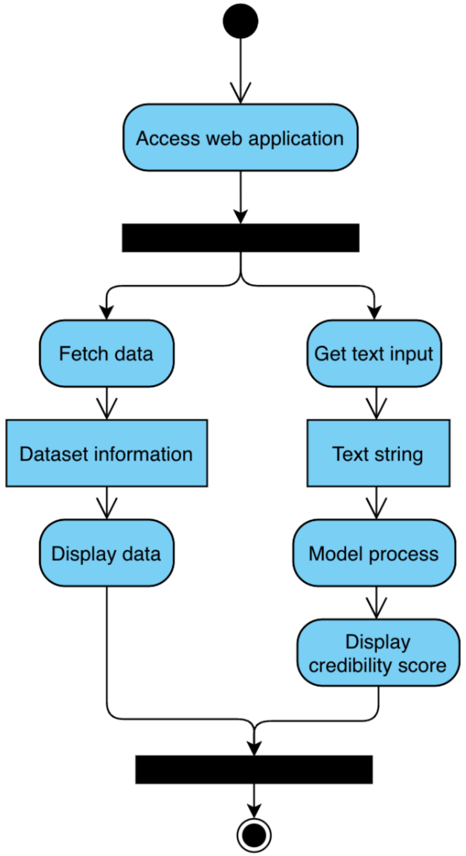
全球流行趨勢
كاس مصر
1.14萬 推文

Web Dashboard



Analysis

Detection



Access web application

Fetch data

Get text input

Dataset information

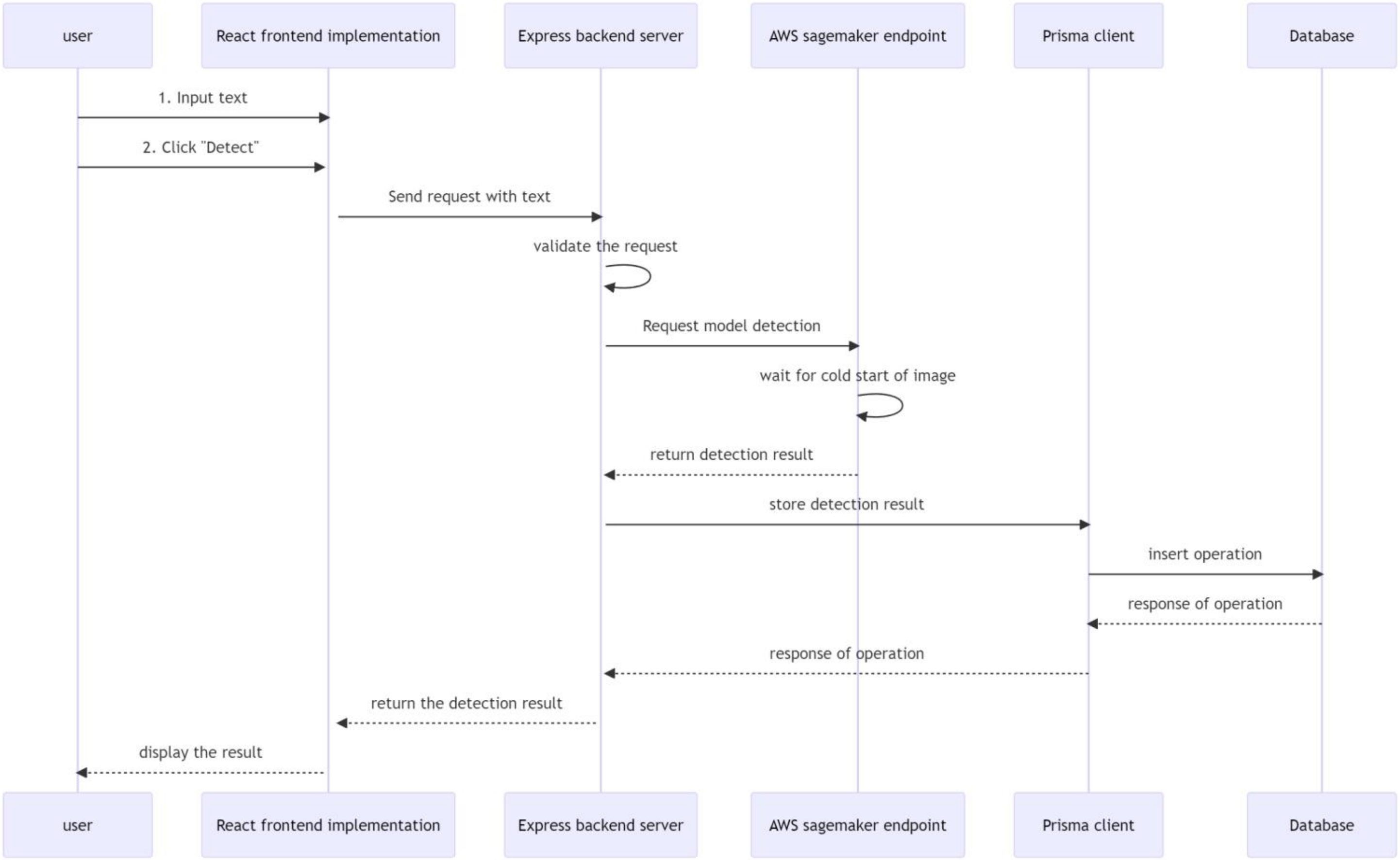
Text string

Display data

Model process

Display credibility score

Web Dashboard Detection



user

React frontend implementation

Express backend server

AWS sagemaker endpoint

Prisma client

Database

1. Input text

2. Click "Detect"

Send request with text

validate the request

Request model detection

wait for cold start of image

return detection result

store detection result

insert operation

response of operation

response of operation

return the detection result

display the result

user

React frontend implementation

Express backend server

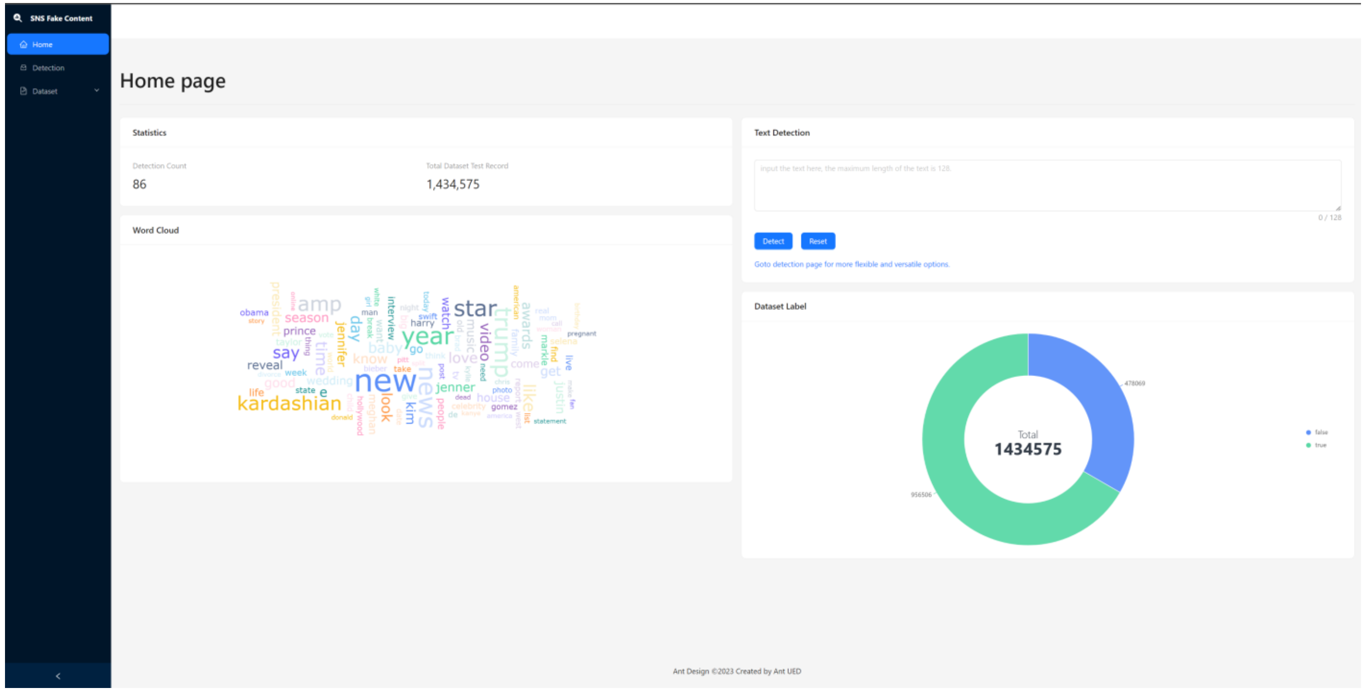
AWS sagemaker endpoint

Prisma client

Database

Web Dashboard – Home Screen

- Home Page
  - Overview of Word Cloud
  - Perform Detection



Home page

Statistics

Detection Count
86

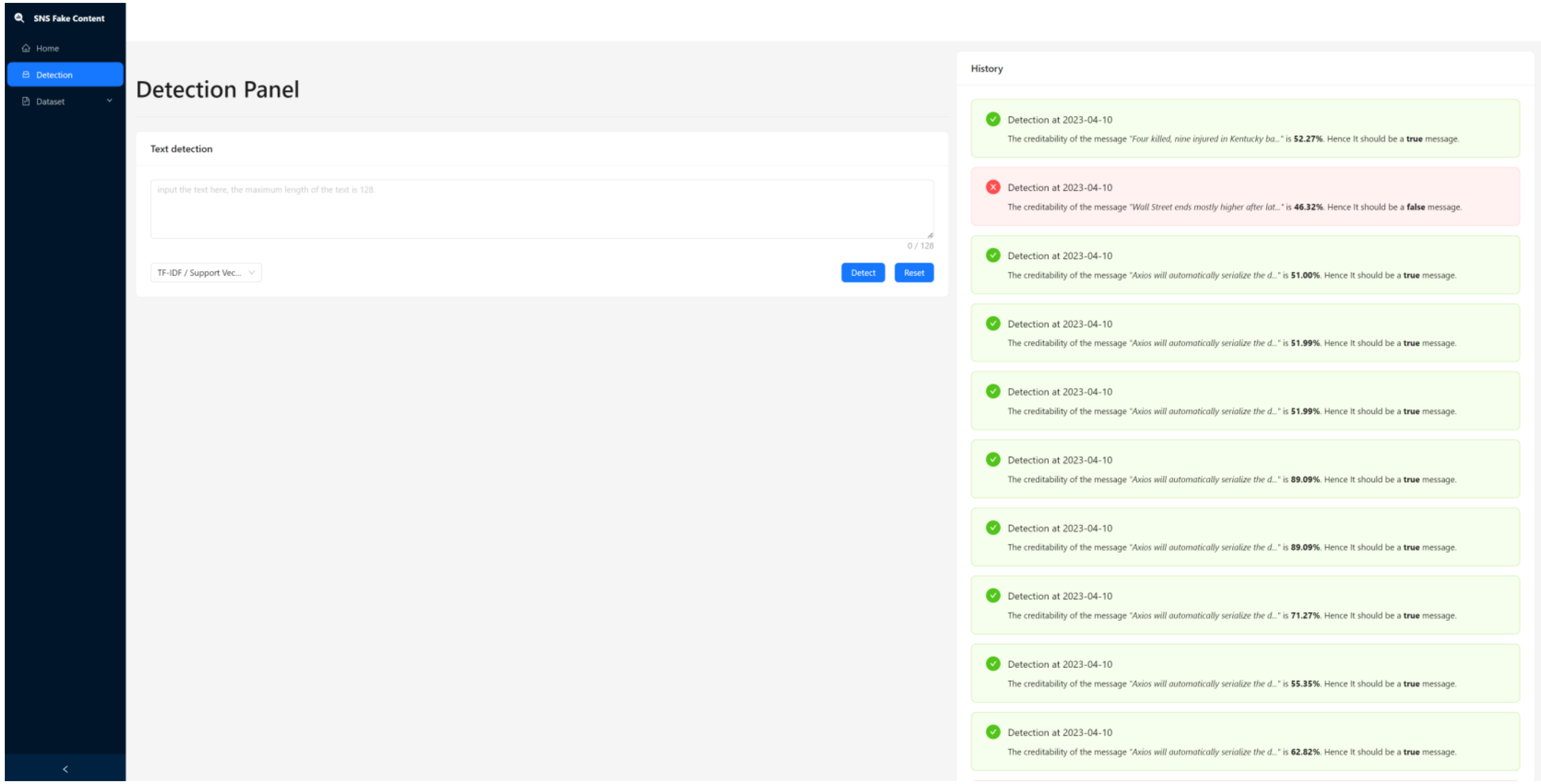
Total Dataset Test Record
1,434,575

input the text here, the maximum length of the text is 128.

0 / 128

Detect
Reset
Goto detection page for more flexible and versatile options.
Dataset Label
Total
1434575
478069
956506
false
true
Ant Design ©2023 Created by Ant UED

Web Dashboard – Detection



Detection Panel

Text detection

input the text here, the maximum length of the text is 128.

0 / 128

TF-IDF / Support Vec...

History

✓ Detection at 2023-04-10
The creditability of the message *"Four killed, nine injured in Kentucky ba..."* is **52.27%**. Hence It should be a **true** message.

✕ Detection at 2023-04-10
The creditability of the message *"Wall Street ends mostly higher after lat..."* is **46.32%**. Hence It should be a **false** message.

✓ Detection at 2023-04-10
The creditability of the message *"Axios will automatically serialize the d..."* is **51.00%**. Hence It should be a **true** message.

✓ Detection at 2023-04-10
The creditability of the message *"Axios will automatically serialize the d..."* is **51.99%**. Hence It should be a **true** message.

✓ Detection at 2023-04-10
The creditability of the message *"Axios will automatically serialize the d..."* is **51.99%**. Hence It should be a **true** message.

✓ Detection at 2023-04-10
The creditability of the message *"Axios will automatically serialize the d..."* is **89.09%**. Hence It should be a **true** message.

✓ Detection at 2023-04-10
The creditability of the message *"Axios will automatically serialize the d..."* is **89.09%**. Hence It should be a **true** message.

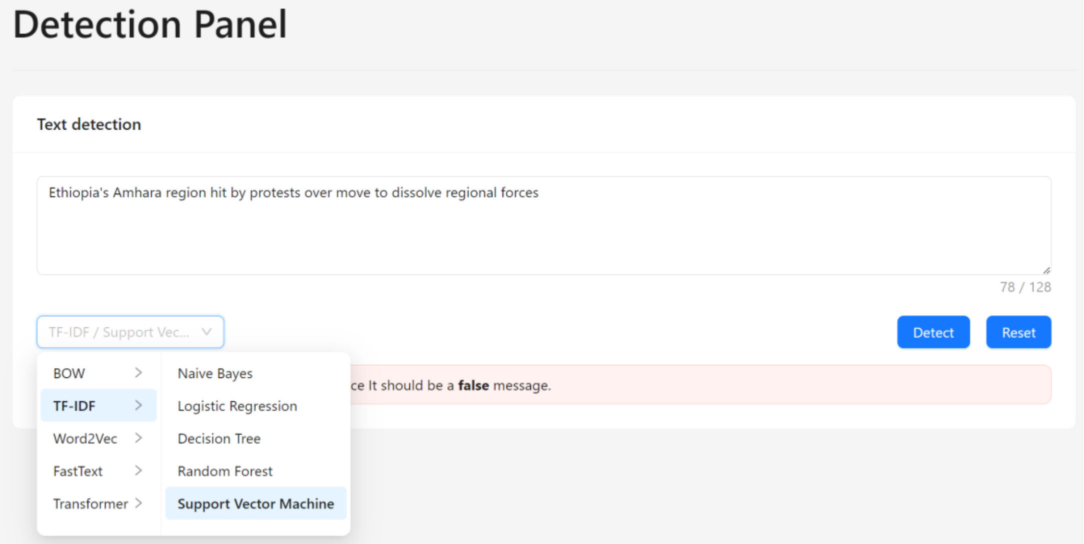
✓ Detection at 2023-04-10
The creditability of the message *"Axios will automatically serialize the d..."* is **71.27%**. Hence It should be a **true** message.

✓ Detection at 2023-04-10
The creditability of the message *"Axios will automatically serialize the d..."* is **55.35%**. Hence It should be a **true** message.

✓ Detection at 2023-04-10
The creditability of the message *"Axios will automatically serialize the d..."* is **62.82%**. Hence It should be a **true** message.

Web Dashboard – Detection

Detection Textbox



Detection Panel

Text detection

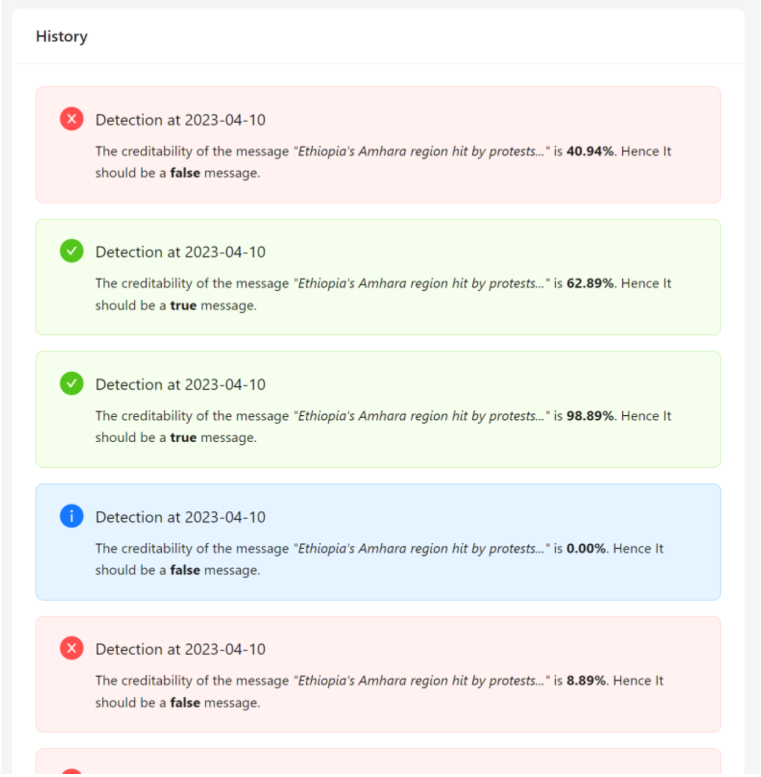
Ethiopia's Amhara region hit by protests over move to dissolve regional forces

78 / 128

TF-IDF / Support Vec...

ce It should be a **false** message.

Detection History



History

Detection at 2023-04-10

The creditability of the message "*Ethiopia's Amhara region hit by protests...*" is **40.94%**. Hence It should be a **false** message.

Detection at 2023-04-10

The creditability of the message "*Ethiopia's Amhara region hit by protests...*" is **62.89%**. Hence It should be a **true** message.

Detection at 2023-04-10

The creditability of the message "*Ethiopia's Amhara region hit by protests...*" is **98.89%**. Hence It should be a **true** message.

Detection at 2023-04-10

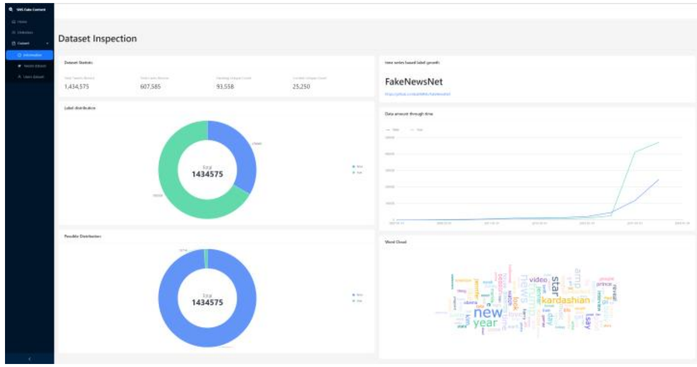
The creditability of the message "*Ethiopia's Amhara region hit by protests...*" is **0.00%**. Hence It should be a **false** message.

Detection at 2023-04-10

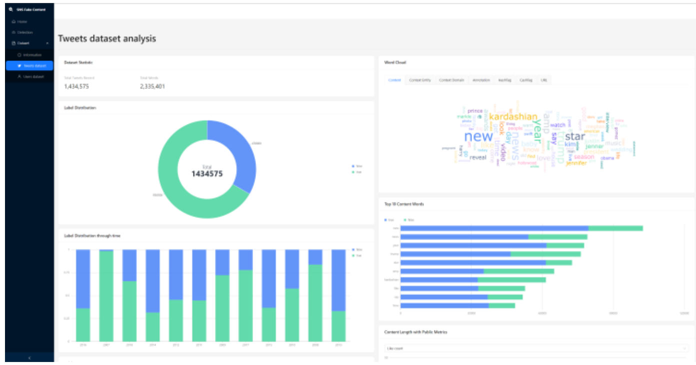
The creditability of the message "*Ethiopia's Amhara region hit by protests...*" is **8.89%**. Hence It should be a **false** message.

Web Dashboard – Dataset Analysis

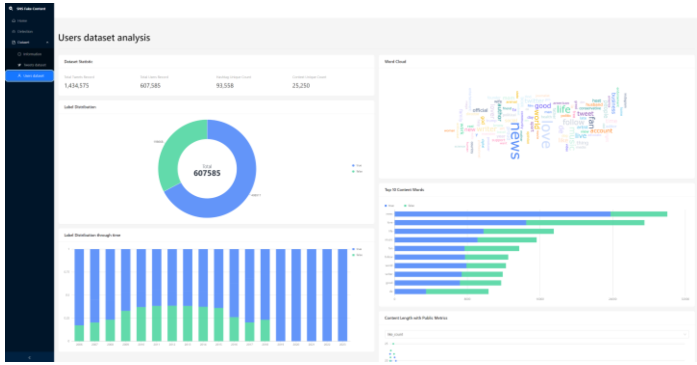
Dataset Information



Tweets' Information

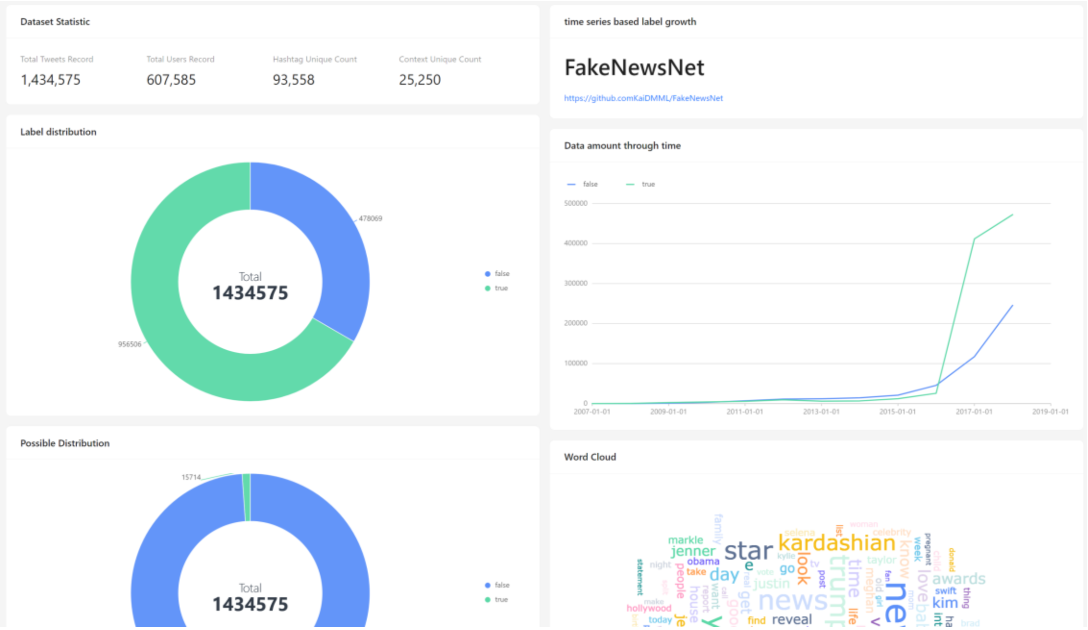


Users' Information



Web Dashboard – Dataset Information

- General Dataset Information
  - Label Distribution
  - Possible Sensitive Content
  - Data Amount through time
  - Word Cloud



Dataset Statistic

Total Tweets Record 1,434,575

Total Users Record 607,585

Hashtag Unique Count 93,558

Context Unique Count 25,250

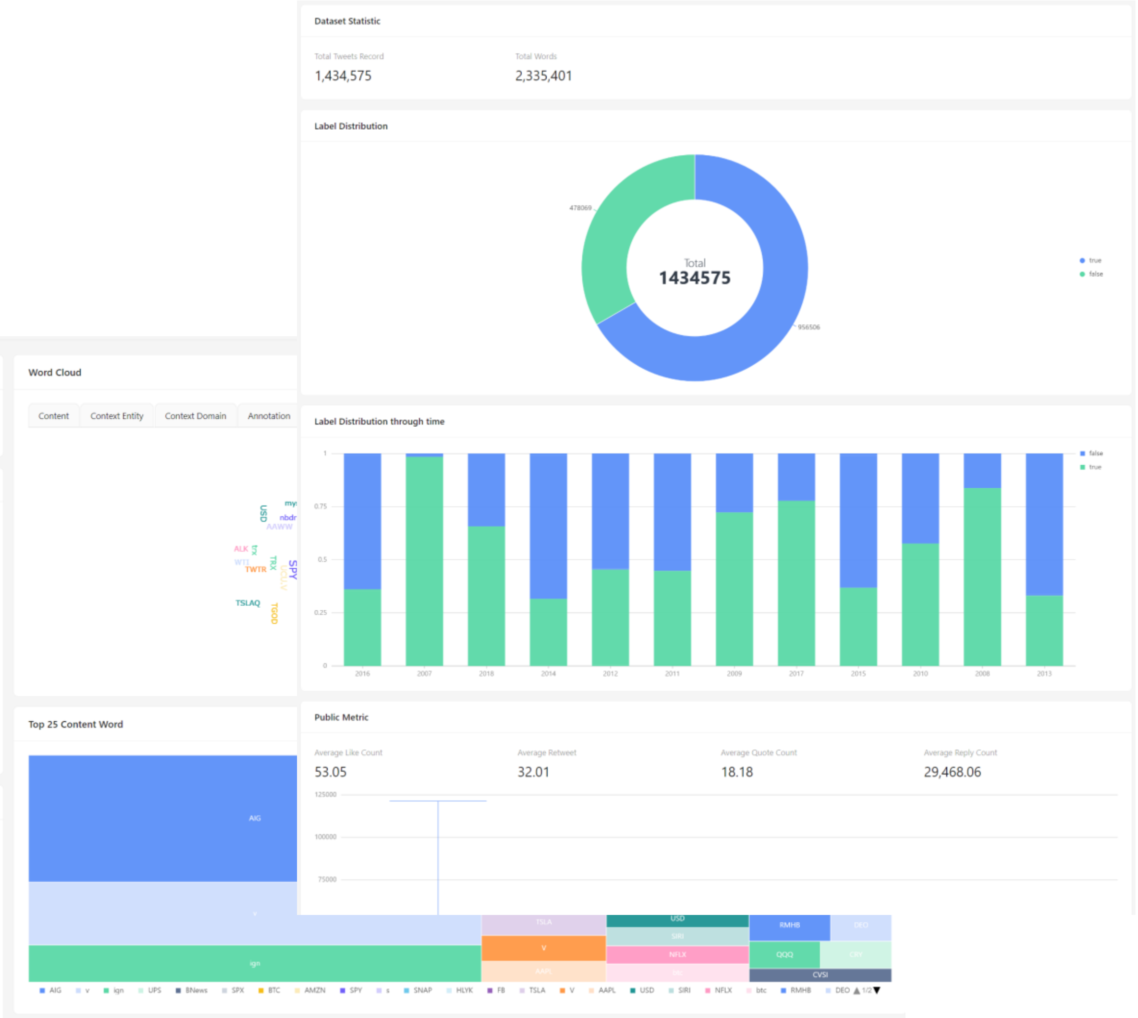
time series based label growth

FakeNewsNet

https://github.comKaiDMML/FakeNewsNet

Web Dashboard – Tweets Information

- Statistic
- Label Distribution
  - All
  - Through Time
- Public Metrics
- Term Frequency
  - Content
  - Entity
  - Context Domain
  - Annotation
  - Hashtag
  - Cashtag
  - Url



Dataset Statistic

Total Tweets Record: 1,434,575

Total Words: 2,335,401

Public Metric

Average Like Count: 53.05

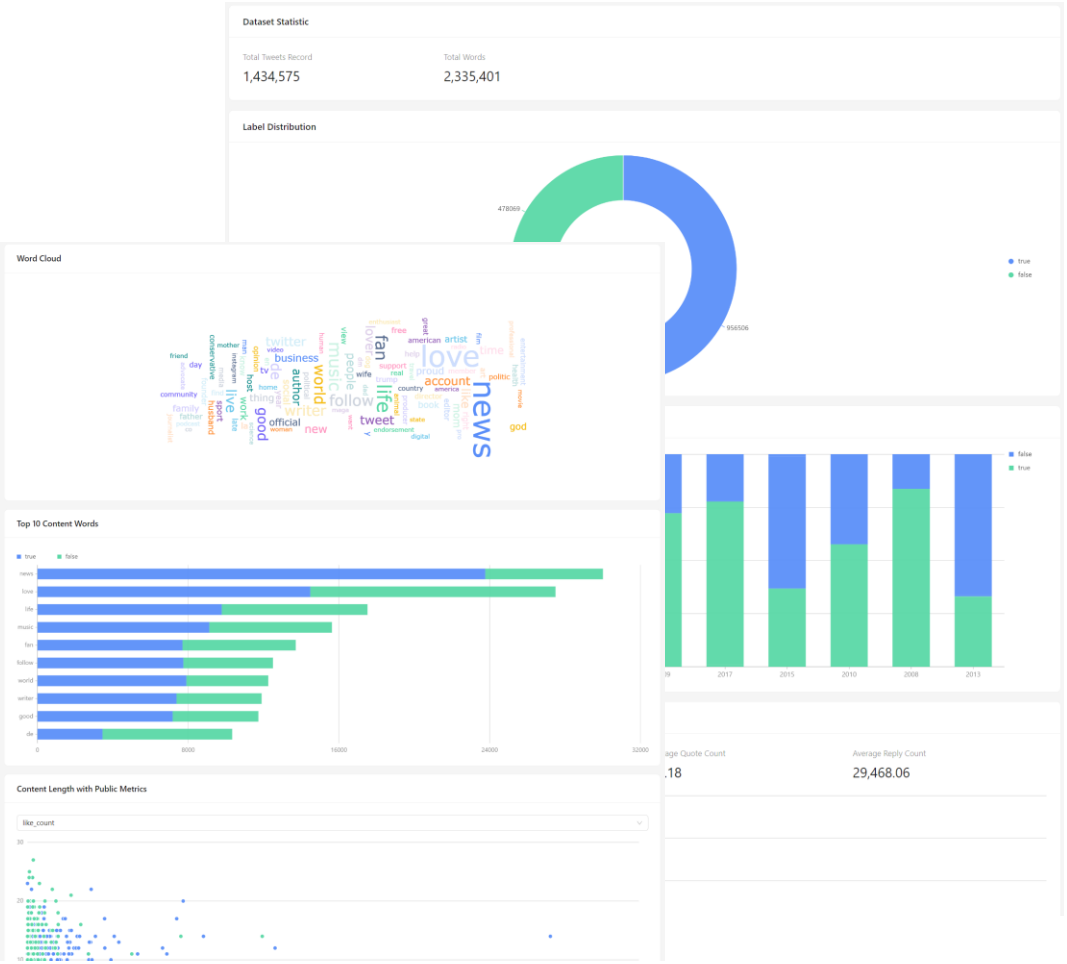
Average Retweet: 32.01

Average Quote Count: 18.18

Average Reply Count: 29,468.06

Web Dashboard – Users Information

- Data Statistic
- User Description information



Dataset Statistic

Total Tweets Record
1,434,575

Total Words
2,335,401

Label Distribution

Word Cloud

Top 10 Content Words

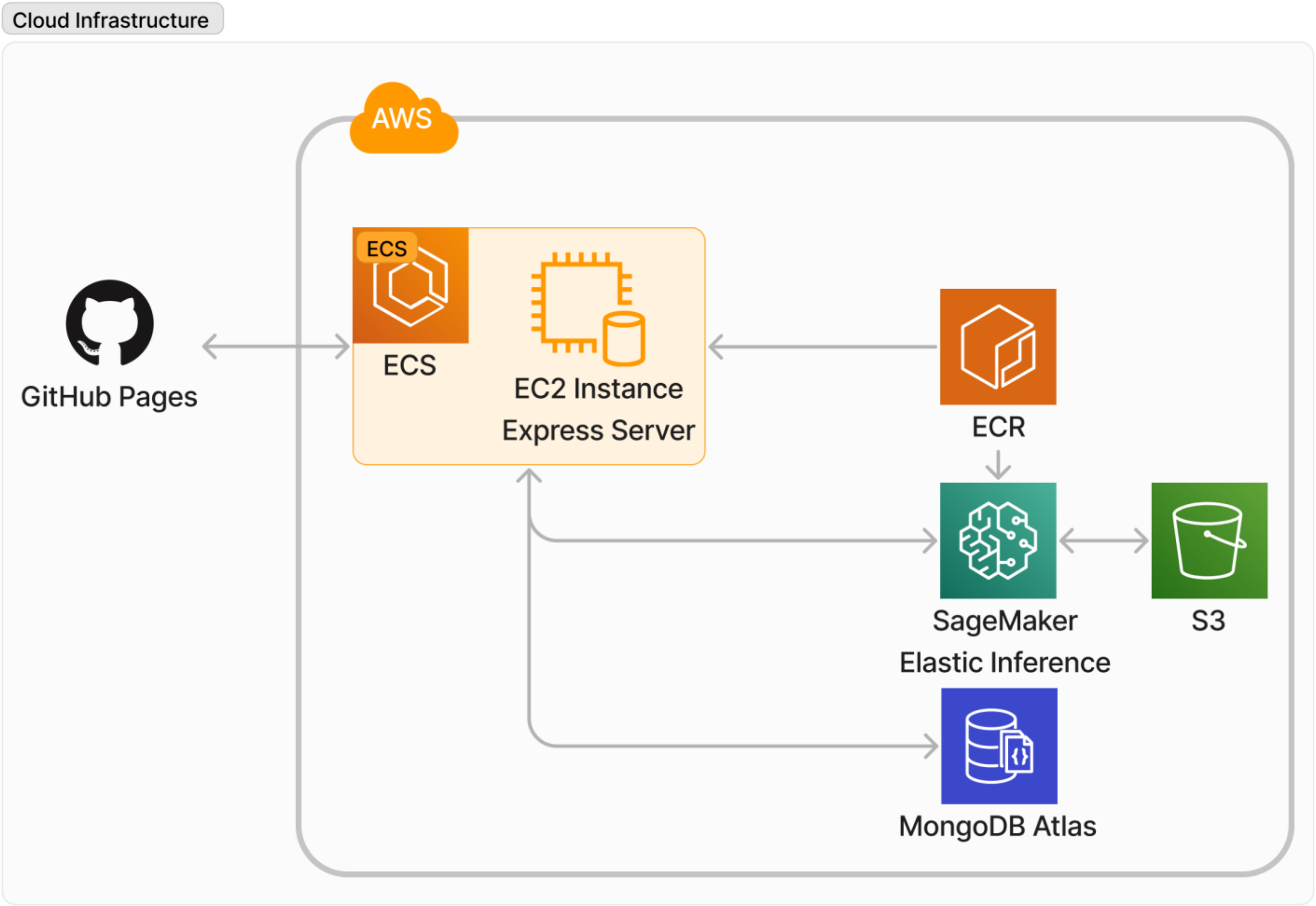
Content Length with Public Metrics

like_count

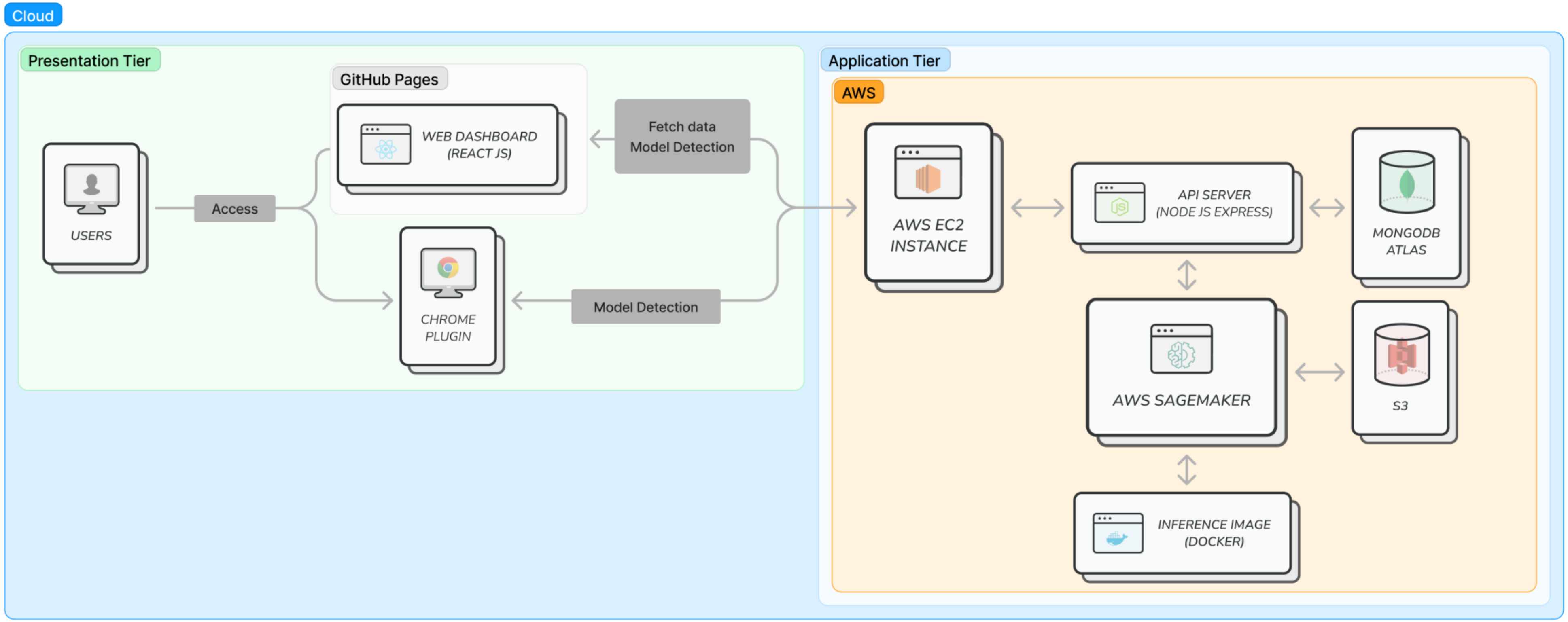
…age Quote Count
.18

Average Reply Count
29,468.06

Deployment


Cloud Infrastructure
AWS
GitHub Pages
ECS
ECS
EC2 Instance
Express Server
ECR
SageMaker
Elastic Inference
S3
MongoDB Atlas

Deployment


Cloud
Presentation Tier
USERS
Access
GitHub Pages
WEB DASHBOARD (REACT JS)
Fetch data Model Detection
CHROME PLUGIN
Model Detection
Application Tier
AWS
AWS EC2 INSTANCE
API SERVER (NODE JS EXPRESS)
MONGODB ATLAS
AWS SAGEMAKER
S3
INFERENCE IMAGE (DOCKER)

Conclusion

Future work and Improvement

Analysis

- More detail on analysis
- Unsupervised learning on feature
  - Clustering

Model

- Expand dataset

Application

- Optimize detection performance
- Real time analysis
  - Scrap social media content
  - Perform detection and analysis in data pipeline

Conclusion

Classification Model

- Different model algorithm
- Different token size, data size

Implementation

- Web dashboard
- Chrome Plugin

The End

Q&A Section